2019

Clustering of temporal gene expression data with mixtures of mixed effects models

BOSTON UNIVERSITY

GRADUATE SCHOOL OF ARTS AND SCIENCES

Dissertation

CLUSTERING TEMPORAL GENE EXPRESSION DATA WITH A PENALIZED MIXTURE OF MIXED EFFECTS MODELS

by

DARLENE LU

M.S., University of Wisconsin, Madison, 2009
B.S., University of Washington, Seattle, 2006

Submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

2019

Approved by

First Reader _____

Serkalem Demissie, PhD
Associate Professor of Biostatistics

Second Reader _____

Yorghos Tripodis, PhD
Associate Professor of Biostatistics

Third Reader _____

Louis Gerstenfeld, PhD
Professor of Orthopaedic Surgery

Fourth Reader _____

Howard Cabral, PhD
Professor of Biostatistics

DEDICATION

This dissertation is dedicated to my father, who has always pushed me to aim for the stars.

I would like to thank my mentor and advisor, Serkalem Demissie, who has patiently provided me with an endless supply of guidance and wisdom. I would not be the Statistician I am today without it!

To my other committee members, Yorghos Tripodis, Howard Cabral, Louis Gerstenfeld, and Kerrie Nelson, thank you for all your support and help through this process. Thank you Yorghos for always pointing me in the right direction after coming to you with a crazy new idea. Dr. G, I am grateful to you for all the context and biological insight you have provided. Your enthusiasm and excitement for my work has always helped me push through the lows. Howard, thank you for your constant words of encouragement and support.

I would also like to thank my family, who continually believed in me and pushed me to be my best. Sherman, without you, my happiness throughout this endeavor would not have been so bright. Thank you for your patience and support throughout these years.

To my BU classmates, it has been amazing getting to know each and every one of you and this experience would not have been the same without you guys. I am excited to see what the future has in store for all of us.

CLUSTERING TEMPORAL GENE EXPRESSION DATA WITH A PENALIZED MIXTURE OF MIXED EFFECTS MODELS

DARLENE LU

Boston University, Graduate School of Arts and Sciences, 2019

Major Professor: Serkalem Demissie, PhD

Associate Professor of Biostatistics

ABSTRACT

While time-dependent processes are important to biological functions, methods to leverage temporal information from large data have remained computationally challenging. In temporal gene-expression data, clustering can be used to identify genes with shared function in complex processes. Algorithms like K-Means and standard Gaussian mixture-models (GMM) fail to account for variability in replicated data or repeated measures over time and require a priori cluster number assumptions, evaluating many cluster numbers to select an optimal result. An improved penalized-GMM offers a computationally-efficient algorithm to simultaneously optimize cluster number and labels.

The work presented in this dissertation was motivated by mice bone-fracture models interested in determining patterns of temporal gene-expression during bone-healing progression. To solve this, an extension to the penalized-GMM was proposed to account for correlation between replicated data and repeated measures over time by introducing random-effects using a mixture of mixed-effects polynomial regression models and an entropy-penalized EM-Algorithm (EPEM).

First, performance of EPEM for different mixed-effects models were assessed

with simulation studies and applied to the fracture-healing study. Second, modifications to address the high computational cost of EPEM were considered that either clustered subsets of data determined by predicted polynomial-order (S-EPEM) or used modified-initialization to decrease the initial burden (I-EPEM). Each was compared to EPEM and applied to the fracture-healing study. Lastly, as varied rates of fracture-healing were observed for mice with different genetic-backgrounds (strains), a new analysis strategy was proposed to compare patterns of temporal gene-expression between different mice-strains and assessed with simulation studies. Expression-profiles for each strain were treated as separate objects to cluster in order to determine genes clustered into different groups across strain.

We found that the addition of random-effects decreased accuracy of predicted cluster labels compared to K-Means, GMM, and fixed-effects EPEM. Polynomial-order optimization with BIC performed with highest accuracy, and optimization on subspaces obtained with singular-value-decomposition performed well. Computation time for S-EPEM was much reduced with a slight decrease in accuracy. I-EPEM was comparable to EPEM with similar accuracy and decrease in computation time. Application of the new analysis strategy on fracture-healing data identified several distinct temporal gene-expression patterns for the different strains.

CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF SYMBOLS AND ABBREVIATIONS

AIC	.....	Akaike's Information Criterion
BIC	.....	Bayesian Information Criterion
CI	.....	Confidence Interval
Cov	.....	Covariance
CV	.....	Cross-validation
$\mathbb{E}$	.....	Expectation
EM	.....	Expectation-Maximization
EPEM	.....	Entropy Penalized EM Algorithm
FE	.....	Fixed Effects
GMM	.....	Gaussian Mixture Models
LOOCV	.....	Leave-one-out cross-validation
ME	.....	Mixed Effects
MLE	.....	Maximum Likelihood Estimation
p	.....	gene-specific fixed effects polynomial order
pdf	.....	Probability Distribution Function
q	.....	gene-specific random effects polynomial order
$\mathbb{R}$	.....	the Real plane
RE	.....	Random Effects

ρ	.....	Correlation
SVD	.....	Singular Value Decomposition
Var	.....	Variance

CHAPTER 1

Introduction

1.1 BACKGROUND AND MOTIVATIONS

Cluster analysis is a method of grouping objects into a set of distinct and disjoint classes, called clusters. Objects within a cluster are very similar, whereas objects in other clusters are more dissimilar. Many clustering algorithms have proven to be useful in understanding gene function, gene regulation, and cellular processes. Genes that have similar expression patterns (co-expressed genes) would be grouped together as they would potentially share similar cellular functions and be involved in the same cellular processes. Cluster analysis can be used to obtain these unobserved groups leading to new information on the functions of certain genes that are not currently known (Eisen et al., 1998).

In recent times, the decreasing costs of microarray experiments have allowed investigators to conduct even more elaborate designs, where measurements are taken over several time-points. Microarray experiments are subject to a degree of variability in the measurements, therefore, biological replicates using replicated arrays are often needed to improve the stability of gene expression estimation. Therefore, clustering algorithms should be able to account for the correlation between repeated measurements over time or replicated data.

The work in this dissertation is motived by several prior microarray gene expression studies that aimed to investigate the impacts of genetic variability on the post-natal regenerative processes of fracture healing (Wigner et al., 2010) (Grimes et al., 2011). A bone fracture was introduced into the femoral bone from three strains of mice (with different genetic backgrounds). At carefully selected time

points during the bone healing process, microarray gene-expression measurements were extracted from the fracture site. Three replicates of fracture calluses were obtained at each of the ten time points from day 0 through day 35. Correlation exists between repeated measurements from the same gene over time and biological replicates at each time point. Ignoring the covariance structure between the repeated measurements or biological replicates could result in a failure to capture important sources of variability leading to incorrect inferences or groupings of the temporal gene expression profiles. More details regarding the bone fracture-healing experiment is discussed in Appendix A.1.

1.2 COMMON CLUSTERING ALGORITHMS

One commonly used clustering algorithm is called K-means (Hartigan & Wong, 1979). Given a set of observations $(x_1, x_2, ..., x_n)$, where each observation is a d-dimensional vector of gene-expression measurements for gene i, K-means tries to partition the n observations into K sets $(C_1, C_2, ..., C_K)$, where the number of sets (K) is pre-specified, by optimizing the following objective function

$$\sum_{k=1}^{K} \sum_{x \in C_k} ||x - \mu_k||^2, \tag{1.1}$$

x is a temporal gene-expression profile in cluster C_k, and μ_k is the centroid (mean of the objects) of C_i. In other words, the objective function aims to minimize the sum of squared distances of all temporal profiles from their cluster center. The benefit of K-means clustering is that it is simple and very fast. However, one drawback is that the number of gene clusters, K, is not usually known. To determine the optimal number of clusters, the algorithm must be run multiple times with differ-

ent values of K to compare cluster results. Additionally, the additional variability from biological replicates and repeated measurements over time of the temporal gene expression data is not accounted for in the K-means clustering, which may lead to incorrect clusters.

In contrast to K-means, a partition-based clustering algorithm, hierarchical clustering generates a hierarchical series of nested clusters that can be represented by a tree (dendrogram) (Sneath & Sokal, 1973). The branches of the dendrogram has information on the nested structured grouping of the data as well as how similar clusters are to one another. The number of clusters can be obtained by cutting the dendrogram at a specific level. Hierarchical clustering can be divided into two general approaches, an agglomerative (bottom-up) approach and a divisive (top-down) approach. Agglomerative algorithms initialize each temporal gene expression profile as its own cluster and merges each pair of clusters until all groups are merged into one cluster. Divisive algorithms does the opposite. They start with one cluster that includes all temporal gene expression profiles and split the data at each step until distinct clusters of temporal profiles remain. Different measures of cluster proximity are used (Kaufman & Rousseeuw, 2009). However, hierarchical clustering algorithms suffer from a high computational complexity (Jain et al., 1999), and a lack of refinement of previous clustering's (once a decision has been made, good or bad, it cannot be undone in the following steps).

An alternative approach involves model-based clustering (Banfield & Raftery, 1993). In the family of model-based clustering, a model based approach assumes that data are generated from a mixture model of probability distributions. Model-based clustering in microarray analysis, with normalized gene expression values, typically utilizes a Gaussian mixture model (GMM), where each cluster is gener-

ated from a normal probability density function. In gene-expression data, Yeung et al. (Yeung et al., 2001) found that the commonly used data transformations were sufficient to satisfy normality assumptions. The Expectation Maximization (EM) Algorithm is used to estimate the mixture model parameters, as well as the missing cluster label for each gene (Dempster et al., 1977). However, the EM algorithm is sensitive to chosen initializations resulting in optimization issues (Biernacki & Celeux, 2003). Similar to K-means, the number of components (K) to use in the clustering algorithm must also be determined, which adds an additional complexity to the problem.

To account for the temporal nature of the data, modifications to the Gaussian mixture model have been proposed from using mixtures of fixed effects polynomial regression models (Gaffney & Smyth, 1999) to mixed effects regression models (Celeux et al., 2005) (Ng et al., 2006). Additionally, to solve the component number optimization problem, a penalty term was added to the likelihood function to create a data-driven algorithm to simultaneously optimize component number and the mixture model parameters (Yang et al., 2012) (Chamroukhi, 2015). However, the penalty term was only explored in the context of fixed effects models. If sufficient heterogeneity is expected within a cluster, the addition of random effects to the regression equation could result in clusters with lower misclassification error rates. Additionally, due to the high dimensionality of microarray data, where measurements are taken from tens-of thousands of genes at a time, this clustering process is computationally intensive. To the best of my knowledge, no modifications have been explored to decrease the computational burden of the algorithm.

1.3 DISSERTATION OUTLINE

In **Chapter 2**, we introduce an extension of the algorithm proposed in (Chamroukhi, 2015) that allows for between- and within-replicate variability with mixtures of mixed-effects polynomial regression models to cluster temporal gene expression profiles. Different underlying mixed-effects regression models will be considered (by varying the number of random-effects) and compared to the models proposed in (Yang et al., 2012) and (Chamroukhi, 2015). With simulation studies, the new entropy-penalized EM algorithm (EPEM) will be compared to K-Means and GMM, which both use a two-step approach using BIC to optimize component number. The algorithm proposed here assumes a p-th order mixed-effects polynomial regression equation to approximate the temporal-expression profile for all genes. The clustering algorithm is applied to the motivating example (fracture healing data) accounting for strain-specific variability in the mixed-effects model.

In **Chapter 3**, to counteract the computational inefficiency in high-dimensional data, two modified versions of EPEM will be proposed: (1) a split clustering algorithm, which clusters data in subsets of temporal expression profiles with the same predicted polynomial order and (2) a modified initialization approach that pre-groups the genes into expression profiles with similar combinations of polynomial regression coefficients. Performances of each method will be compared to EPEM with a simulation study. Additionally, the consequence of the choice of p on the misclassification error (MCE) of the cluster labels will be assessed. We will consider models where p is over-specified and under-specified based on simulated data. Finally, model-selection techniques using AIC, BIC or K-fold cross-validation

will be used to determine the optimal p to use in the mixed effects polynomial regression equation. The modified algorithm is applied to the fracture healing data accounting for strain-specific variability in the mixed-effects model and compared to the results in Chapter 2.

In **Chapter 4**, we apply the methods developed in the previous chapters to the fracture healing data. One goal of the study was to identify strain-specific differences on patterns of gene-expression as bone-healing progresses. A new analysis strategy with the EPEM clustering algorithm was proposed to do this. A simulation study was conducted to determine performance of this strategy for pre-defined scenarios. In this analysis strategy, expression-profiles for each strain were treated as separate objects to cluster in order to determine genes clustered into different groups across strain. Parametric and non-parametric pairwise comparisons of measures such as the under the curve (AUC), time to maximum or minimum expression (Tmax or Tmin), and maximum or minimum expression over the bone healing process (Emax or Emin) are conducted on the previously identified gene sets to determine if vertical or horizontal shifts occurred. Finally, an enrichment-analysis using KEGG pathways (Kyoto Encyclopedia of Genes and Genomes (KEGG) Pathways;
http://www.genome.jp/kegg/ pathway.html) is conducted on sets of genes that exhibited significant horizontal or vertical shifts.

CHAPTER 2

Entropy penalized mixture of mixed effects regression models

2.1 BACKGROUND

Model-based clustering has become a popular method in cluster analysis of microarray data ((Ng et al., 2006); (McLachlan et al., 2002); (Celeux et al., 2005); (Chamroukhi, 2015) are a few). It provides a mathematical structure to the data and the clustering procedure by assuming the data to be generated from a mixture of probability distributions. In microarray studies, it is commonly assumed that each component of the mixture model is defined by a multivariate normal density, and by using the properties of normal densities, easily derived solutions can be obtained (McLachlan et al., 2002). However, extensions to mixture models defined in the exponential family of distributions can also be obtained (Lindsay, 1986).

In non-model based approaches, (i.e. K-Means (Hartigan & Wong, 1979) and hierarchical clustering (Sneath & Sokal, 1973)), the temporal ordering or variability between repeated measurements at each time-point of the same gene are ignored. However, in model-based clustering, with the help of polynomial regression equations, we can now account for the ordering and variability between time-points of the same gene. Plenty of work has been done in this field. Examples include (1) a Gaussian mixture model (GMM) allowing each cluster to be represented by its own mean vector and covariance matrix (Fraley et al., 2012), (2) mixtures of fixed effects (FE) polynomial regression models (PRM) for temporal data (an extension to the GMM where each cluster is represented by parameters defined by a polynomial curve (Gaffney & Smyth, 1999)), and (3) a further extension to mixtures of mixed effects (ME) linear regression models (Celeux et al., 2005) (Ng et al., 2006) to

account for correlation between biological replicates or between genes in temporal or cross-sectional data.

A popular method to estimate the mixture parameters is by maximum likelihood estimation (MLE) using the Expectation-Maximization (EM) algorithm (Dempster et al., 1977). However, the EM algorithm is prone to optimization issues resulting from its sensitivity to initialization values (Biernacki & Celeux, 2003). In previous studies, the ME model was found to perform better than GMM when variability from biological replicates was present in the data (Celeux et al., 2005). However, with an unknown cluster number, component number optimization is an additional step to the process. A common solution to this problem is to use model selection criteria such as Akaike's Information Criterion (AIC) (Aike, 1974), Bayesian Information Criterion (BIC) (Fraley et al., 2012) (Roeder et al., 1997) or Integrated Complete Likelihood (ICL) (Biernacki et al., 2000) (McLachlan & Peel, 2000). While shown to perform well in estimating the number of components (Fonesca & G.M.S, 2007), these methods can be computationally intensive, especially in high-dimensional data with a large number of clusters or genes (Fraley et al., 2012).

To counteract the computational inefficiency, a penalty term was proposed to be added to the objective function allowing the algorithm to learn at each iteration and discard illegitimate clusters (Yang et al., 2012). We refer to this as the Entropy Penalized EM (EPEM) algorithm, which can be used to obtain the optimal component size and class labels of each gene in one-step. However, it was only formulated to be applied to multivariate data with no assumptions on the functional relationships in temporal data. Chamroukhi (Chamroukhi, 2015) extended the EPEM to apply it to temporal data by fitting mixtures of FE polynomial or

spline regression models. The algorithm was compared to K-means and the GMM and found that his algorithm performed much better with a lower misclassification error rate (MCE). However, considerations for high-dimensional data or data with multiple replicates per gene were not assessed.

By using a ME mixture model (Celeux et al., 2005) and entropy penalized maximum likelihood estimation (Yang et al., 2012), we propose an extension to the clustering algorithm (Chamroukhi, 2015) for temporal data with additional sources of variability (from repeated measurements over time and biological replicates) and an unknown component number. Different underlying ME polynomial regression models (PRMs) will be considered and compared to the models proposed in (Yang et al., 2012) and (Chamroukhi, 2015). The new EPEM algorithm will also be compared to K-Means and GMM, which both use a two-step approach using BIC to optimize component number.

2.2 FINITE MIXTURE MODELS

Assume that the data consist of N temporal gene-expression profiles. Each profile consists of gene-expression measurements obtained at T time points, which is represented by y_i where i = 1, . . . , N. Now assume that for each time-point we have R replicates so that each temporal gene-expression profile has T$\times$R measurements. The goal is to cluster the N temporal gene-expression profiles into K groups using Gaussian finite mixture models.

First, let us assume that R=1 resulting in T observations for each gene (so we can drop the r-th index). Gaussian finite mixture models assume each temporal gene-expression profile is generated from a mixture of Multivariate Normal distributions, where the k-th cluster (component) has a probability of occurrence,

$\alpha_k = P(z_i = k)$, such that $\sum_{k=1}^{K} \alpha_k = 1$, where z_i is the missing class label (unknown cluster membership) for the temporal profile of gene i. A probability density function can be defined for y_i assuming that the expression values for gene i originated from cluster k.

If each cluster is allowed to have its own set of parameters and probability density function, a set of parameters can be defined such that $\Theta = (\theta_1, \ldots, \theta_K, \alpha_1, \ldots, \alpha_K)$ where $\theta_k = (\beta_k, \sigma_k^2)$ are the cluster-specific parameters for the k-th cluster ($k = 1, \ldots, K$). The overall density for one realization of the data can be represented as a weighted sum of these component densities. It follows that the mixture density for y_i can be represented as (2.1), and it can be shown that the log-likelihood can be formulated as (2.2).

$$p(y_i|\Theta) = \sum_{k=1}^{K} \alpha_k p_k(y_i|\theta_k) \tag{2.1}$$

$$l(\Theta) = logL(\Theta) = \sum_{i=1}^{N} log\{\sum_{k=1}^{K} \alpha_k p_k(y_i|\theta_k)\} \tag{2.2}$$

Maximizing the log-likelihood function (2.2) is not trivial and a closed-form solution does not exist. However, if the problem can be formulated in a missing data framework, the EM algorithm can be used by assuming the cluster memberships (z_i) are unknown. Recall that z_i represents the class label for temporal profile for gene i, and $z_i = c(z_{i1}, \ldots, z_i)$ where $z_{ik} = \mathbb{I}\{y_i \text{originated from cluster k}\}$. The complete data vector is (y_i, z_i), therefore, it can be shown that the complete-data log-likelihood can be formulated as (2.3).

$$l_c(\Theta) = \sum_{i=1}^{N} \sum_{k=1}^{K} z_{ik} log\alpha_k + \sum_{i=1}^{N} \sum_{k=1}^{K} z_{ik} logp_k(y_i|\theta_k) \tag{2.3}$$

The EM-algorithm (Dempster et al., 1977) can then be used to determine the maximum likelihood estimates (MLE) of the parameters. The usefulness of this formulation is that, at each iteration, we are now able to obtain closed-form solutions of our parameters. If x_i is a T-dimensional vector of time-points from which each gene expression measurement is obtained, a cluster-specific conditional probability model, $p_k(y_i|x_i, \theta_k)$, can be defined that associates y_i to x_i using a polynomial regression equation (i.e. $y_i = \beta_0 + \beta_1 x_i + \ldots + \beta_p x_i^p + \epsilon_i$ (Gaffney & Smyth, 1999)). Assuming a normal distribution, $(y_i|x_i)$ is now defined with the probability density function from a multivariate normal distribution with mean $x_i\beta_k$ and covariance Σ_k. The problem becomes clustering a mixture of polynomial regression equations that can be solved via the EM algorithm to obtain probabilities of cluster memberships.

2.3 MIXED-EFFECTS MODEL

Now assume that for each time-point we have R replicates, where each temporal gene-expression profile has T×R measurements. We want to account for the variability between the replicated arrays at each time point and allow the temporal trend for each gene to deviate from its cluster-specific temporal trend, yet still exhibit the same underlying pattern that distinguishes that cluster from the rest. Naturally, a mixed effects model does just that when the genes within a cluster are allowed to have their own intercept and/or slope. The mixture model discussed in **Section 2.2** can be extended to include random effects. The general form of the mixed effects model (2.4) includes a replicate-specific random effect as well as time-varying gene-specific random effects, where p is the order of the polynomial for the fixed effects (FE) and q is the order of the polynomial for the gene-specific

random effects (REs).

$$\begin{aligned} y_{trik} &= \beta_{0k} + \beta_{1k}x_{trik} + ... + \beta_{pk}x_{trik}^{p} \\ &\quad + b_{0ik} + b_{1ik}x_{trik} + \ldots + b_{qik}x_{trik}^{q} + c_{rik} + \epsilon_{trik} \end{aligned} \tag{2.4}$$

$\beta_k = (\beta_{0k}, \beta_{1k}, \ldots, \beta_{pk})$ is the k-th cluster FE coefficient; x_{tri} is t-th time-point for the r-th replicate of the i-th temporal gene-expression profile; $b_{ik} = (b_{0ik}, b_{1ik}, b_{qik})^T \sim N(0, G_k)$ is the gene-specific RE; $c_{rik} \sim N(0, \tau_k^2)$ is the replicate-specific RE (between replicate variability); $\epsilon_{trik} \sim N(0, \sigma_k^2)$ is the measurement error (within-replicate variability). Assume that b_{ik}, c_{rik} and ϵ_{trik} are mutually independent and each cluster is allowed to have a different mean vector and covariance matrix.

In this dissertation, we consider the following mixed-effects models:

- Eq0r1: $y_{trik} = \beta_{0k} + \beta_{1k}x_{trik} + ... + \beta_{pk}x_{trik}^{p} + b_{0ik} + c_{rik} + \epsilon_{trik}$
- Eq1r1: $y_{trik} = \beta_{0k} + \beta_{1k}x_{trik} + ... + \beta_{pk}x_{trik}^{p} + b_{0ik} + b_{1ik}x_{trik} + c_{rik} + \epsilon_{trik}$
- Eq2r1: $y_{trik} = \beta_{0k} + \beta_{1k}x_{trik} + ... + \beta_{pk}x_{trik}^{p} + b_{0ik} + b_{1ik}x_{trik} + b_{2ik}x_{trik}^{2} + c_{rik} + \epsilon_{trik}$
- Eq2r0: $y_{tik} = \beta_{0k} + \beta_{1k}x_{tik} + ... + \beta_{pk}x_{tik}^{p} + b_{0ik} + b_{1ik}x_{tik} + b_{2ik}x_{tik}^{2} + \epsilon_{tik}$

The mixed effect models (Eq0r1, Eq1r1 and Eq2r1) considered estimate $\beta_{0k}, \beta_{1k}, \ldots, \beta_{pk}$, σ_k^2, G_k, τ_k^2, where $G_k \in \mathbb{R}^{(q+1)\times(q+1)}$ and q=0, 1 and 2, respectively. Lastly, Eq2r0 model estimates $\beta_{0k}, \beta_{1k}, \ldots, \beta_{pk}$, σ_k^2, G_k, where $G_k \in \mathbb{R}^{3\times3}$, which ignores the variability of the replicates. These models differ by the number of random effects that are allowed.

The REs can be viewed as unobserved missing data and the EM algorithm (Dempster et al., 1977) (McLachlan & Peel, 2000) can be used to determine the maximum likelihood parameter estimates by iteratively maximizing the conditional

expectation of the complete data log-likelihood (2.5),

$$\begin{aligned} l_c(\Theta) &= logL(\Theta|y,b,c,z) \\ &= log\prod_{i=1}^{N}\prod_{k=1}^{K}[\alpha_k p_k(y_i|b_i,c_i,\theta_k)p_k(b_i|\theta_k)p_k(c_i|\theta_k)]^{z_{ik}} \\ &= \sum_{i=1}^{N}\sum_{k=1}^{K} z_{ik}log\alpha_k + \sum_{i=1}^{N}\sum_{k=1}^{K} z_{ik}log[p_k(y_i|b_i,c_i,\theta_k)] \\ &+ \sum_{i=1}^{N}\sum_{k=1}^{K} z_{ik}log[p_k(b_i|\theta_k)] + \sum_{i=1}^{N}\sum_{k=1}^{K} z_{ik}log[p_k(c_i|\theta_k)] \end{aligned} \tag{2.5}$$

where $p_k(y_i|b_i,c_i,\theta_k)$, $p_k(b_i|\theta_k)$, and $p_k(c_i|\theta_k)$ are conditional probability density functions, which are defined in the next section.

2.4 ENTROPY PENALIZED EM-ALGORITHM (EPEM)

Several problems still exist with the standard EM mixture model clustering algorithm. Initialization of the EM algorithm is important, which can yield different results when initialized at different locations of the parameter space (Biernacki & Celeux, 2003). To solve this, many algorithms require multiple random initializations and choosing the results for the one with the largest log-likelihood or using initializations from other clustering methods such as K-Means (Figueiredo et al., 2002). However, this can be computationally intensive, particularly for large datasets. Additionally, the number of components in the standard EM needs to be specified. To choose the optimal number of components, many algorithms adapt a brute-force method by repeatedly clustering the data assuming a range of cluster numbers and selecting the optimal clustering based on some criterion. This can be time consuming and limited to the range of cluster sizes considered.

To avoid this, we propose to add a penalty term (Yang et al., 2012) in the same

spirit as (Chamroukhi, 2015) to the standard log-likelihood function (2.5). The number of clusters in our data can be represented by the unknown class labels (z_i) for each temporal gene expression profile. As the number of clusters in our data increase making the model more complex, the entropy also increases. Using the entropy of z_i, $H(z_i)$, as the penalty term in our penalized log-likelihood maximization function allows a penalty to occur for overly complex models. The derivation of $H(z_i)$ and $H(z)$ is shown below, where $z = (z_1, \ldots, z_N)$.

$$\begin{aligned} H(z_i) &= E[-log(P(z_i))] \\ &= -\sum_{k=1}^{K} P(z_i = k) log[P(z_i = k)] = -\sum_{k=1}^{K} \alpha_k log \alpha_k \\ H(z) &= E[-log(P(z))] \\ &= -\sum_{i=1}^{N}\sum_{k=1}^{K} \alpha_k log \alpha_k = -N \sum_{k=1}^{K} \alpha_k log \alpha_k \end{aligned} \tag{2.6}$$

Adding the penalty term, $H(z)$ to (2.5), the new penalized log-likelihood (2.7) is obtained.

$$l_{c(p)}(\Theta) = l_c(\Theta) - \lambda H(z) \tag{2.7}$$

Maximum Likelihood Estimation

The probability distributions can be defined for each of the components of (2.5) or (2.7) with normally distributed probability distribution functions; $(y_i|b_i, c_i, \theta_k) \sim N(X\beta_k + Ub_{ik} + Vc_{ik}; \sigma_k^2 I_{TR})$; $(b_{ik}|\theta_k) \sim N(0, G_k)$; $(c_{ik}|\theta_k) \sim N(0, \tau_k^2 I_{TR})$, where X is the Vandermonde matrix of time-points of order p, U is a Vandermonde matrix of time-points of order q, and V is a diagonal matrix of size TR×R where the di-

agonals are unit vectors of length R. A simplified description of the EM Algorithm is presented here, see **Appendix A.2** on the derivation of the penalized likelihood function and **Appendix A.3** and **Appendix A.4** for derivations of components of the Expectation and Maximization step of the EM-Algorithm.

E-step

The E-step (Expectation step) of the EM algorithm takes the expectation of the complete data-log likelihood (2.7) conditional on the observed data and the parameter space of the previous iteration, resulting in the following objective function:

$$\begin{aligned} J(\Theta, \Theta^{(s)}) = & \sum_{i=1}^{N}\sum_{k=1}^{K} w_{ik}^{(s)} log\alpha_k \\ & + \sum_{i=1}^{N}\sum_{k=1}^{K} w_{ik}^{(s)} \left\{ -\frac{TRlog\sigma_k^2}{2} - \frac{\mathbb{E}[D_i' D_i | y_i, \Theta^{(s)}]}{2\sigma_k^2} \right\} \\ & + \sum_{i=1}^{N}\sum_{k=1}^{K} w_{ik}^{(s)} \left\{ -\frac{log|G_k|}{2} - \frac{\mathbb{E}[b_{ik}' G_k^{-1} b_{ik} | y_i, \Theta^{(s)}]}{2} \right\} \\ & + \sum_{i=1}^{N}\sum_{k=1}^{K} w_{ik}^{(s)} \left\{ -\frac{Rlog\tau_k^2}{2} - \frac{\mathbb{E}[c_{ik}' c_{ik} | y_i, \Theta^{(s)}]}{2\tau_k^2} \right\} \\ & + \lambda \sum_{i=1}^{N}\sum_{k=1}^{K} \alpha_k log\alpha_k \end{aligned} \tag{2.8}$$

where $D_i = y_i - X\beta_k - Ub_{ik} - Vc_{ik}$ and $w_{ik}^{(s)}$ is the posterior probability that the temporal gene-expression profile belongs to the k-th cluster conditioning on the observed data and the current parameter estimates $\Theta^{(s)}$. The updating equation for $w_{ik}^{(s)}$ is given in (2.9)

$$\begin{aligned}\hat{w}_{ik}^{(s)} &= \mathbb{E}[z_{ik}|y_i, \Theta^{(s)}] \\ &= \frac{P(y_i|z_i = k, \Theta^{(s)})P(z_i = k)}{P(y_i|\Theta^{(s)})} \\ &= \frac{\alpha_k^{(s)} p_k(y_i|\theta_k^{(s)})}{\sum_{k=1}^{K} \alpha_k^{(s)} p_k(y_i|\theta_k^{(s)})}\end{aligned} \tag{2.9}$$

To complete the E-step, conditional expectations ($E[b_{ik}|y_i]$, $E[c_{ik}|y_i]$, $E[b_{ik}b_{ik}^{'}|y_i]$ and $E[c_{ik}^{'}c_{ik}|y_i]$) need to be determined (**Appendix A.3**).

M-step

The M-step involves maximizing (2.8) with respect to each of the parameters $(\alpha_k, \beta_k, \sigma_k^2, \tau_k^2, G_k)$ by taking the derivative with respect to each parameter. The formulas for the MLEs of each parameter is given in (2.10) (to solve for α_k, we must use a Lagrange multiplier to satisfy the constraint that $\sum_{k=1}^{K} \alpha_k = 1$). See **Appendix A.4** for more details on derivations of parameters. Note that $(Y_i|X\beta_k, U, V) \sim N(X\beta_k, \Gamma_k)$, where $\Gamma_k = UGU^{'} + \tau_k^2 VV^{'} + \sigma_k^2 I_{TR}$.

$$
\begin{aligned}
\hat{\beta}_k^{(s+1)} =& \frac{\sum_{i=1}^{N} w_{ik}^{(s)}[\hat{\beta}_k^{(s)} + \hat{\sigma}_k^{2(s)}(X^TX)^{-1}X'\Gamma_k^{-1(s)}(y_i - X\hat{\beta}_k^{(s)})]}{W_i} \\
\hat{\tau}_k^{2(s+1)} =& \frac{1}{RW_i}\sum_{i=1}^{N} w_{ik}^{(s)}[\hat{\tau}_k^{4(s)}(y_i - X\hat{\beta}_k^{(s)})^T\Gamma_k^{-1(s)}VV'\Gamma_k^{-1(s)} \\
& (y_i - X\hat{\beta}_k^{(s)}) + R\hat{\tau}_k^{2(s)} - \hat{\tau}_k^{4(s)}tr(\Gamma_k^{-1(s)}VV')] \\
\hat{G}_k^{(s+1)} =& \frac{1}{W_i}\sum_{i=1}^{N} w_{ik}^{(s)}[\hat{G}_k^{(s)}U'\Gamma^{-1(s)}(y_i - X\hat{\beta}_k^{(s)})(y_i - X\hat{\beta}_k^{(s)})' \\
& \Gamma_k^{-1(s)}U\hat{G}_k^{(s)} + \hat{G}_k^{(s)} - \hat{G}_k^{(s)}U'\Gamma_k^{-1(s)}UG_k^{(s)}] \\
\hat{\sigma}_k^{2(s+1)} =& \frac{1}{TRW_i}\sum_{i=1}^{N} w_{ik}^{(s)}[\hat{\sigma}_k^{4(s)}(y_i - X\hat{\beta}_k^{(s)})'\Gamma_k^{-1(s)}\Gamma_k^{-1(s)} \\
& (y_i - X\hat{\beta}_k^{(s)}) + tr(Cov(D_{ik}|y_i, \Theta^{(s)})] \\
\hat{\alpha}_k^{(s+1)} =& \frac{W_i}{N} + \lambda\hat{\alpha}_k^{(s)}\left\{log\hat{\alpha}_k^{(s)} - \sum_{k=1}^{K}\hat{\alpha}_k^{(s)}log\hat{\alpha}_k^{(s)}\right\}
\end{aligned}
\tag{2.10}
$$

where $W_i = \sum_{i=1}^{N} w_{ik}^{(s)}$, $\Gamma_k^{-1} = (UG_kU^T + \tau_k^2VV' + \sigma_k^2I_{TR})^{-1}$, and $Cov(D_{ik}|y_i, \Theta^{(s)}) = U[\hat{G}_k^{(s)} - \hat{G}_k^{(s)}U'\Gamma_k^{-1(s)}U\hat{G}_k^{(s)}]U' + V[\hat{\tau}_k^2I_R - \hat{\tau}_k^{4(s)}V'\Gamma_k^{-1(s)}V]V'$.

λ is also estimated in an iterative way. $\lambda \in [0, 1)$ controls how much of a role the penalty term plays in the objective function. If λ is close to zero the penalized-likelihood function is close to the standard likelihood function (2.5) and more stable, exhibiting convergence properties of the standard EM. If λ is large (i.e. close to 1), the penalty term plays a larger role by creating competition between clusters. It forces the algorithm to get rid of illegitimate clusters with small proportions. For cluster k, if $log\hat{\alpha}_k^{(s)} - \sum_{k=1}^{K}\hat{\alpha}_k^{(s)}log\hat{\alpha}_k^{(s)} > 0$, then $H(z)$ is small and the second term is

positive, causing $\hat{\alpha}_k^{(s+1)}$ to increase. However, if $log\hat{\alpha}_k^{(s)} - \sum_{k=1}^{K} \hat{\alpha}_k^{(s)} log\hat{\alpha}_k^{(s)} < 0$, then $\hat{\alpha}_k^{(s+1)}$ will decrease. If it is lower than a set threshold, the cluster will be deemed illegitimate and removed by the algorithm. Note that this penalty term does not guarantee an increased likelihood function at each iteration of the algorithm. However, through simulation studies it was shown that the penalized log-likelihood function levels off after convergence of the algorithm (Chamroukhi, 2015). $\lambda^{(s+1)}$ is determined by $\hat{\alpha}_k^{(s)}$, $\hat{\alpha}_k^{(s+1)}$ and $\hat{w}_{ik}^{(s)}$, defined in (2.11).

$$\lambda^{(s+1)} = min\{\frac{\sum_{k=1}^{K} exp(-\eta n|\alpha_k^{(s+1)} - \alpha_k^{(s)}|)}{K}, \frac{(1 - max_k \frac{\sum_{i=1}^{N} w_{ik}^{(s)}}{n})}{-max_k \alpha_k^{(s-1)} \sum_{k=1}^{K} \alpha_k^{(s)} log\alpha_k^{(s)}}\} \tag{2.11}$$

where η is set to be $\min\{1, 0.5^{\lfloor \frac{T}{2} - 1 \rfloor}\}$. The motivation and derivation of λ is presented in much detail in (Yang et al., 2012) and (Chamroukhi, 2015). λ is updated so that it is large when $\hat{\alpha}_k$ is not changing enough between iterations to promote cluster competition. However, when $\hat{\alpha}_k$ exhibits a large change, λ needs to be small to promote stability of the algorithm.

2.4.1 Algorithm

2.4.1.1 Initialization

To initialize the algorithm, the number of clusters in the mixture model is set to equal the number of temporal gene-expression profiles in the dataset. To obtain cluster-specific parameter estimates, a FE PRM is fit to the temporal-gene expression profiles for each gene to obtain $\beta_k^{(0)}$ (vector of size N) for each cluster k. For all clusters, $\sigma_k^{2(0)}$ and $\tau_k^{2(0)}$ are initialized to be the median residual error from all

Table 2.1: EM-Algorithm for entropy-penalized maximum likelihood estimation

1. **Initialize EM algorithm**: Start with $K^{(0)}$ = N different clusters
a. Compute $w_{ik}^{(0)}$ from $\beta_k^{(0)}$, $\tau_k^{(0)}$, $G_k^{(0)}$, $\sigma_k^{2(0)}$ and $\alpha_k^{(0)} = 1/N$
b. Let $\lambda^{(0)}$=1 and converge=0
2. **For s = 1:**
a. Compute $\alpha_k^{(1)}$ from $w_{ik}^{(0)}$, $\lambda^{(0)}$, and $\alpha_k^{(0)}$
b. Compute $\lambda^{(1)}$ from $\alpha_k^{(1)}$, $\alpha_k^{(0)}$ and $w_{ik}^{(0)}$
c. Update $K^{(1)} = K^{(0)}$ - number of components where $\alpha_k^{(1)} \leq 1/N$
d. Normalize $w_{ik}^{(0)}$ and $\alpha_k^{(1)}$ to satisfy the sum to 1 constraint
e. Compute $\beta_k^{(1)}$, $\tau_k^{2(1)}$, $G_k^{(1)}$, $\sigma_k^{2(1)}$ from $w_{ik}^{(0)}$
f. Compute $w_{ik}^{(1)}$ from $\beta_k^{(1)}$, $\tau_k^{2(1)}$, $G_k^{(1)}$, $\sigma_k^{2(1)}$ and $\alpha_k^{(1)}$
3. **For s $>$ 2:**
a. Compute $\alpha_k^{(s)}$ from $w_{ik}^{(s-1)}$, $\lambda^{(s-1)}$, and $\alpha_k^{(s-1)}$
b. Compute $\lambda^{(s)}$ from $\alpha_k^{(s)}$, $\alpha_k^{(s-1)}$ and $w_{ik}^{(s-1)}$
c. Update $K^{(s)} = K^{(s-1)}$ - # of components where $\alpha_k^{(s)} \leq 1/N$
d. If s $\geq$ 60 and $K^{(s-60)} - K^{(s)} = 0$, then let $\lambda^{(s)} = 0$
e. Normalize $w_{ik}^{(s-1)}$ and $\alpha_k^{(s)}$ to satisfy the sum to 1 constraint
f. Compute $\beta_k^{(s)}$, $\tau_k^{2(s)}$, $G_k^{(s)}$, $\sigma_k^{2(s)}$ from $w_{ik}^{(s-1)}$
h. Compute $w_{ik}^{(s)}$ from $\beta_k^{(s)}$, $\tau_k^{2(s)}$, $G_k^{(s)}$, $\sigma_k^{2(s)}$ and $\alpha_k^{(s)}$
i. If $\max_k \|\beta_k^{(s)} - \beta_k^{(s-1)}\| < \epsilon$ then converge=1 and STOP, else s = s+1 and repeat step 3

gene- or replicate-specific models run, respectively (to avoid singularities when the covariance matrix is too small). $G_k^{(0)}$ is initiated as $\sigma_k^{2(0)} \times (U^{'}U)^{-1}$ resulting in a (q+1)×(q+1) matrix where q is determined by the chosen random-effects polynomial order for EPEM. Finally, $\alpha_k^{(0)} = 1/N$. The posterior probability ($w_{ik}^{(0)}$) is initialized by $\lambda = 0$, $\alpha_k^{(0)}$, $\beta_k^{(0)}$, $\sigma_k^{2(0)}$, $G_k^{(0)}$ and $\tau_k^{2(0)}$. Lastly, the order of the fixed (p) and mixed (q) effects polynomial order must be specified prior to running the algorithm. The same polynomial order is fit to all clusters. Model selection methods such as cross-validation may be used to select the optimal order (Gaffney, 2004), which will be assessed in **Chapter 3**.

2.4.1.2 *Iteration s of the EPEM algorithm*

For iteration s, the maximization step updates the parameters in our model defined in (2.10). At each subsequent iteration, the algorithm alternates between the E- and M-step to remove illegitimate clusters with small $\alpha_k < 1/N$. After stability of cluster number is reached (no clusters are removed for 60 consecutive iterations), λ is set to zero and convergence of the algorithm can proceed and is assessed by the FE coefficients such that $max_k||\beta_k^{(s)} - \beta_k^{(s+1)}|| < \epsilon = 10^{-4}$. At the last iteration, predicted cluster labels ($\hat{z}_i$) is determined by setting membership to be the cluster with the largest $\hat{w}_{ik}$ over all k for each gene ($\hat{z}_i = \text{argmax}_k \hat{w}_{ik}$). The algorithm is described in detail in Table 2.1.

2.5 SIMULATION STUDY

Simulations were performed to assess the performance of each ME considered by measuring the misclassification error (MCE) of each clustering. MCE is defined as the error rate for a given cluster label relative to the known truth. Data was generated with four (uniform, $\alpha_k = 0.25$, and varying, $\alpha_k \in (0.4, 0.3, 0.2, 0.1)$ mixing proportions) or eight (uniform mixing proportions, $\alpha_k = 1/8$) clusters with varying sample sizes (N=200, 500 or 5000), varying replicates (2, 4, 10), and nine different within- and between-replicate variability (σ_k^2, τ_k^2) specifications with varying orders (all pairwise combinations of $\{0.01, 0.1, 0.4\}$) for varying degree of polynomial (Simulation A and Simulation B). Table 2.2 summarizes the parameter estimates used to generate the data for two types of simulated data. Simulation A data consists of four or eight distinct clusters of linear and quadratic temporal gene-expression profiles with 10 time points, low ($\sigma_k^2 = 0.01$ and $\tau_k^2 = 0.01$) and

Table 2.2: Simulation A and B parameters to obtain datasets used in simulation studies with equal mixing proportions for each of the eight clusters considered. Datasets were generated with varying number of genes per cluster (50, 125 and 1250) and replicates (2, 4 or 10) and within- (σ_k^2) or between- (τ_k^2) replicate variability.

	Parameter	**Cluster 1**	**Cluster 2**	**Cluster 3**	**Cluster 4**	**Cluster 5**	**Cluster 6**	**Cluster 7**	**Cluster 8**
Sim A[a]	β_0	0.5	0.1	0.5	-1	2.5	5.0	3.0	4.0
	β_1	–	0.5	-0.1	1.5	–	-0.2	0.4	-1.0
	β_2	–	–	–	-0.1	–	–	–	0.1
	β_3	–	–	–	–	–	–	–	–
	Var(b_0)	$0.1\sigma^2$	$0.02\sigma^2$	$0.1\sigma^2$	$0.2\sigma^2$	$0.5\sigma^2$	$1.0\sigma^2$	$0.6\sigma^2$	$0.8\sigma^2$
	Var(b_1)	–	$0.1\sigma^2$	$0.02\sigma^2$	$0.3\sigma^2$	–	$0.04\sigma^2$	$0.08\sigma^2$	$0.2\sigma^2$
	Var(b_2)	–	–	–	$0.02\sigma^2$	–	–	–	$0.02\sigma^2$
	Var(b_3)	–	–	–	–	–	–	–	—
	σ^2				0.01 or 0.10				
	τ^2				0.01 or 0.10				
Sim B[b]	β_0	-1.05	-2.3	2.03	2.85	2.30	1.40	1.5	-2.900
	β_1	0.41	0.5	-0.58	-0.5	-0.30	-0.05	–	0.510
	β_2	-0.026	-0.02	0.033	0.024	0.01	–	–	-0.014
	β_3	4.00E-4	2.50E-4	-5.03E-4	-3.56E-4	–	–	–	–
	Var(b_0)	0.21	0.46	0.406	0.57	0.21	0.08	0.09	0.34
	Var(b_1)	0.02	0.03	0.03	0.03	2.25E-4	6.25E-6	–	6.50E-4
	Var(b_2)	5.20E-4	4.00E-4	6.60E-4	4.80E-4	1.00E-8	–	–	1.96E-8
	Var(b_3)	8.00E-6	5.00E-6	1.00E-5	7.20E-6	–	–	–	–
	σ^2	0.01	0.02	0.01	0.03	0.1	0.1	0.01	0.01
	τ^2				0.01				

[a] Correlation (ρ) between b_i such that $\rho(b_0, b_1) = -0.5$, $\rho(b0, b2) = 0.4$ and $\rho(b1, b2) = -0.9$

[b] $\rho(b_0, b_1) = -0.7$, $\rho(b0, b2) = 0.6$, $\rho(b1, b2) = -0.8$,$\rho(b0, b3) = 0.4$, $\rho(b1, b3) = -0.6$, $\rho(b2, b3) = 0.9$

high ($\sigma_k^2 = 0.1$ and $\tau_k^2 = 0.1$) error (panel 1 and 2 of Figure 2.1, respectively). Simulation B data consists of more realistic cubic time-curves and 12 time points (panel 3 of Figure 2.1). Var(b_i), correspond to diagonal entries of the G-matrix, whereas covariance between b_i and b_j for ($i \neq j$) is equal to $\rho(b_i, b_j)\sqrt{Var(b_i)}\sqrt{Var(b_j)}$.

To investigate the impact of high-dimensional data with many clusters, data with 30 clusters were generated with linear and quadratic curves (Supplementary Table B.1; N=6000 with 200 genes per cluster; $\alpha_k = 1/30$; 4 replicates; 10 time points with 1 day intervals). The clustering algorithm (EPEM) was run for p=2 and q=2. Robustness of departures from normality was assessed with Simulation A parameters (N=500 with 125 genes per cluster;$\alpha_k = 0.25$; 4 replicates) where the between- and within-replicate variability is drawn from the Student's T-distribution (10 degrees of freedom; to allow for outliers) and the Gamma distribution ($b_{qik} \sim Gamma(3, 10^{q^2})$, $c_{rik} \sim Gamma(3, 5)$, $\epsilon_{trik} \sim Gamma(3, 10)$; to allow for skewness). Figure 2.2 presents a histogram, for cluster one by underlying distribution (Normal, Student's T, and Gamma), of the expression values at day one from a simulated dataset.

1000 datasets were generated for each scenario. To assess the performance of each method, the predicted cluster number ($\hat{K}$) and MCE are reported. The ME models, Eq0r1, Eq1r1, Eq2r1 and Eq2r0, were compared between each other and additionally with the FE only model proposed in (Chamroukhi, 2015) and the multivariate clustering in (Yang et al., 2012). For Yang's, Chamroukhi's and Eq2r0 models, the EPEM clustering algorithm was run on the mean of the measurements over all replicates of each time point (each temporal gene-expression profile has 10 expression measurements). Data were generated from curves defined by (2.4), where $(p, q) \in \{(1, 1), (2, 2), (3, 3)\}$, corresponding to linear, quadratic and cubic

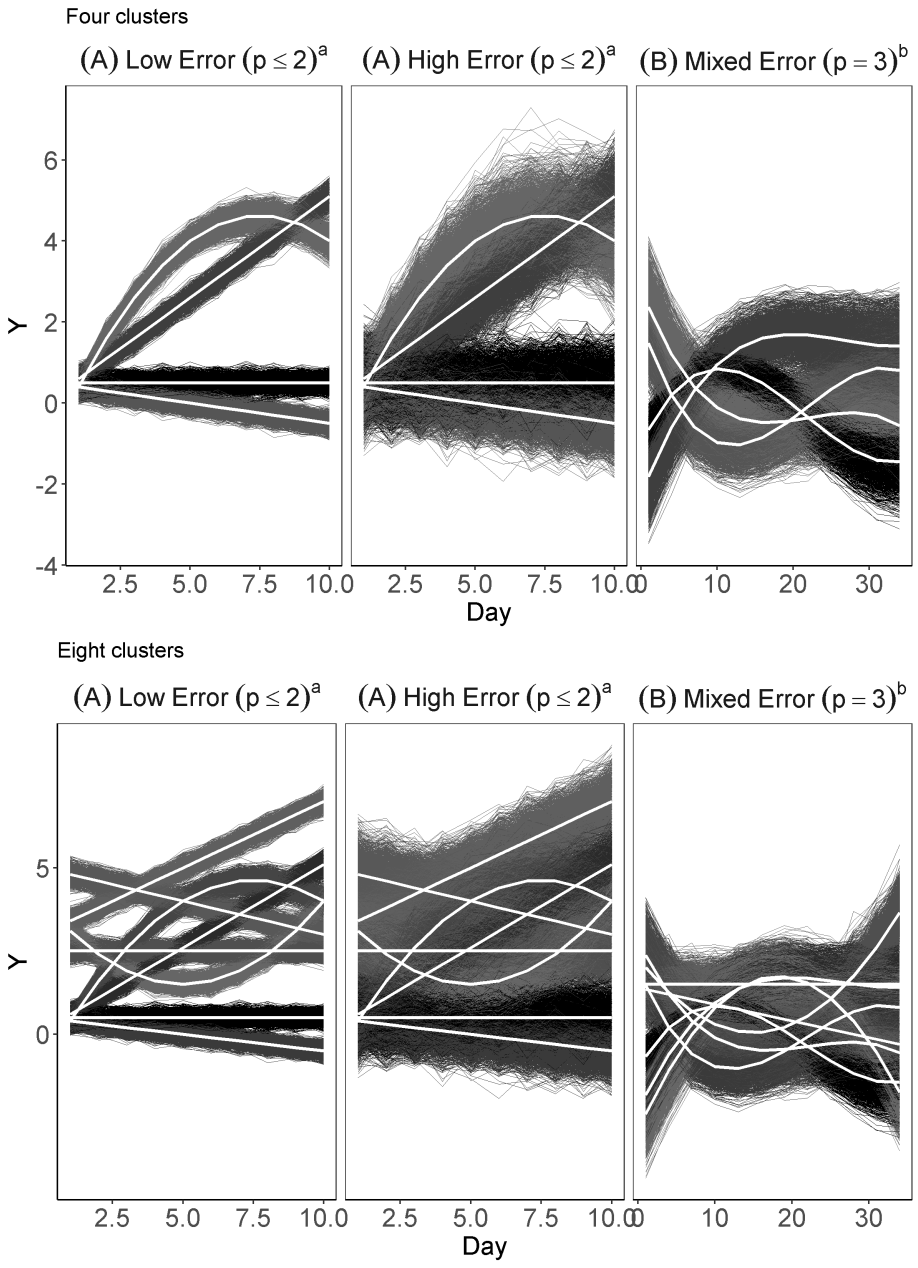

Figure 2.1: One iteration of a dataset generated for simulation studies A and B (4 replicates, 50 genes per cluster and 4 true clusters). ^{a}Low error: σ_k^2=0.01, τ_k^2=0.01; High error: σ_k^2=0.1, τ_k^2=0.1; ^{b} $\sigma_k^2 \in (0.01 - 0.03)$; $\tau_k = 0.01$;

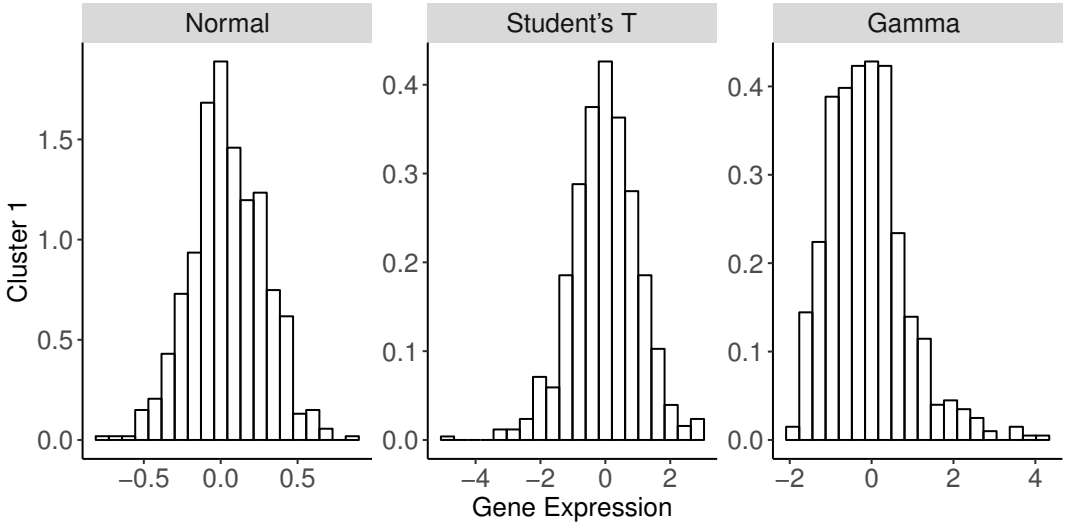

Figure 2.2: Distribution of gene-expression values for one time-point, cluster 1 from Table 2.2, and 1250 genes per cluster. Errors were sampled from a normal distribution with $b_{qik} \sim N(0, Var(b_{qik}))$ from Table 2.2 in the main text, $c_{rik} \sim N(0, 0.01)$, $\epsilon_{trik} \sim N(0, 0.01)$, Student's T-distribution with 10 degrees of freedom or a Gamma where $b_{qik} \sim Gamma(3, 10^{q^2})$, $c_{rik} \sim Gamma(3, 5)$, $\epsilon_{trik} \sim Gamma(3, 10)$

polynomial functions, respectively. The clustering algorithm for simulation A and B was run with p = 2 or 3, respectively, and varying q (0 to 2). $q \leq 2$ was chosen for pragmatic purposes in an attempt to not over-specify the model as model based clustering requires the estimation of a large number of parameters.

The EPEM algorithm with ME models were compared with two easily implemented methods in R, K-Means clustering and GMM using the Stats and MClust package in R (Fraley et al., 2012), respectively. A two-step approach was necessary for KMeans and GMM, where multiple clusterings must be performed over a range of component sizes (2 to 60) and using BIC to select the best set of class labels. $\hat{K}$, MCE and minutes of run time (Linux Centos 6.6 operating system) were compared between the 3 algorithms.

2.6 APPLICATION TO FRACTURE HEALING STUDY

For illustration purposes, we applied the EPEM-Eq2r1 (p=4; q=2) clustering algorithm to the fracture healing study. For each strain and time-point, the data were averaged over the 3 replicates. The replicate-specific random effect, c_{rik}, is included in the model to account for strain-specific variability, as the three strains have different rates of healing (Jepsen et al., 2008). The underlying model for Eq2r1 is shown below in (2.12). The expression values were standardized by the gene-specific mean and standard deviation over all strains. Additional information on the fracture-healing study are given in **Appendix A.1**.

$$y_{ik} = \beta_{0k} + \beta_{1k}x_i + \beta_{2k}x_i^2 + \beta_{3k}x_i^3 + \beta_{4k}x_{tri}^4 + b_{0ik} + b_{1ik}x_i + b_{2ik}x_i^2 + c_{rik} + \epsilon_{ik} \quad (2.12)$$

Table 2.3: EPEM clustering results with different underlying models (Yang, Chamroukhi or EPEM with varying q; Simulation A data with 4 clusters and 4 replicates and 200, 500 or 5000 genes, i.e. 50, 125 or 1250 genes per cluster). Average (SD) predicted cluster number ($\hat{K}$) and misclassification error (MCE %) over 1000 simulated datasets.

			Mean (SD)					
			50 Genes/Cluster		**125 Genes/Cluster**		**1250 Genes/Cluster**	
Data	**Error**[1]	**Model**	$\hat{K}$	**MCE,%**	$\hat{K}$	**MCE,%**	$\hat{K}$	**MCE,%**
Sim A	Low	EPEM-Yang[2]	6.3 (0.8)	24.5 (8.7)	8.3 (0.9)	41.3 (6.4)	15.4 (1.5)	59.5 (2.4)
		EPEM-Chamroukhi (p=2)[3]	6.7 (1.1)	28.1 (10.5)	8.6 (1.0)	43.3 (6.2)	15.3 (1.5)	59.2 (2.9)
		EPEM-Eq0r1 (p=2, q=0)	5.2 (0.7)	12.6 (6.7)	6.0 (0.5)	19.5 (4.3)	6.9 (0.8)	22.9 (1.7)
		EPEM-Eq1r1 (p=2, q=1)	4.3 (0.5)	2.7 (4.7)	4.2 (0.4)	2.1 (4.2)	4.4 (0.5)	4.7 (5.3)
		EPEM-Eq2r1 (p=2, q=2)	4.2 (0.5)	2.2 (4.4)	4.1 (0.3)	0.8 (2.7)	4.2 (0.4)	2.2 (4.2)
		EPEM-Eq2r0 (p=2, q=2, NR)[4]	4.3 (0.5)	2.7 (5.3)	4.0 (0.2)	0.2 (1.7)	4.1 (0.3)	0.7 (2.7)
		KMeans	3.0 (0.0)	25.0 (0.0)	10.0 (0.0)	31.0 (1.6)	11.0 (0.0)	32.5 (0.7)
		GMM	4.2 (0.4)	1.7 (3.6)	4.8 (0.7)	7.4 (6.1)	7.8 (0.4)	20.8 (1.2)
Sim A	High	EPEM-Yang[2]	6.8 (1.1)	30.3 (10.4)	8.7 (1.0)	44.5 (5.9)	15.4 (1.5)	59.8 (2.4)
		EPEM-Chamroukhi (p=2)[3]	6.7 (1.1)	29.5 (10.3)	8.6 (1.0)	44.0 (5.9)	15.4 (1.6)	59.7 (2.8)
		EPEM-Eq0r1 (p=2, q=0)	5.2 (0.7)	13.0 (7.0)	6.0 (0.5)	19.8 (4.2)	7.0 (0.8)	23.0 (1.7)
		EPEM-Eq1r1 (p=2, q=1)	4.3 (0.5)	2.7 (4.7)	4.2 (0.4)	2.0 (4.1)	4.5 (0.5)	5.0 (5.5)
		EPEM-Eq2r1 (p=2, q=2)	4.3 (0.5)	2.6 (4.5)	4.1 (0.3)	1.0 (3.0)	4.2 (0.4)	1.9 (3.9)
		EPEM-Eq2r0 (p=2, q=2, NR)[4]	4.3 (0.5)	3.2 (6.0)	4.0 (0.2)	0.6 (3.4)	4.1 (0.3)	0.9 (3.0)
		KMeans	3.0 (0.2)	26.1 (0.8)	9.0 (0.1)	36.8 (4.2)	16.8 (5.3)	58.8 (12.6)
		GMM	3.5 (0.7)	15.6 (11.4)	4.8 (0.6)	8.4 (6.0)	8.1 (1.6)	22.9 (2.1)
Sim B	Mixed	EPEM-Yang[2]	6.3 (1.4)	26.6 (12.1)	8.2 (1.1)	43.3 (7.3)	33.6 (2.9)	84.3 (1.4)
		EPEM-Chamroukhi (p=3)[3]	5.9 (1.1)	23.3 (11.2)	7.9 (0.8)	41.3 (6.9)	33.5 (3.2)	84.4 (1.4)
		EPEM-Eq0r1 (p=3, q=0)	6.6 (1.3)	28.8 (12.7)	8.3 (0.8)	44.4 (5.1)	20.5(1.8)	69.6 (2.3)
		EPEM-Eq1r1 (p=3, q=1)	4.8 (0.7)	9.9 (8.1)	5.9 (0.8)	19.8 (7.9)	9.0(1.2)	40.1 (7.4)
		EPEM-Eq2r1 (p=3, q=2)	4.4 (0.6)	4.7 (6.4)	4.6 (0.8)	4.5 (6.0)	4.4 (0.7)	2.3 (4.7)
		EPEM-Eq2r0 (p=3, q=2, NR)[4]	4.3 (0.5)	3.0 (5.4)	4.2 (0.4)	1.2 (3.1)	4.0 (0.1)	0.2 (1.3)
		KMeans	6.2 (2.1)	24.8 (4.9)	8.7 (0.8)	36.4 (8.2)	24.2 (0.8)	73.4 (1.7)
		GMM	4.0 (0.1)	0.0 (0.4)	4.0 (0.0)	0.0 (0.01)	5.4 (0.6)	14.6 (6.0)

EPEM=Entropy Penalized EM Algorithm; FE: Fixed Effect; RE: Random Effect;
p: GMM: Standard Gaussian Mixture Model; FE polynomial order; q: RE polynomial order.
[1]Low: σ_k^2=0.01,τ_k^2=0.01; High: σ_k^2=0.1,τ_k^2=0.1; Mixed: $\sigma_k^2 \in (0.01 - 0.03)$; $\tau_k = 0.01$
[2]Yang's model did not utilize a regression model (no fixed- or random-effects) (Yang et al., 2012).
[3]Chamroukhi's model only included fixed-effects in the regression model (Chamroukhi, 2015).
[4]NR: No replicate-specific RE drops c_{rik} from (2.4).

2.7 RESULTS

2.7.1 Varying the polynomial order of the regression model for EPEM

The clustering results from Simulation A (linear and quadratic curves) and B (cubic curves) with uniform mixing proportions are shown in Table 2.3 for data with four clusters, a fixed replicate number (R=4), varied number of genes per cluster (50, 125 and 1250), and different error assumptions (Simulation A, low: σ_k^2= 0.01, τ_k^2=0.01; high: σ_k^2=0.1, τ_k^2=0.1; Simulation B, mixed: $\sigma_k^2 \in$ (0.01, 0.02, 0.01, 0.03), τ_k^2=0.01).

In Simulation A, among the set of EPEM models, the ME models with p=2 and q=2, Eq2r1 and Eq2r0 (no replicate specific effect), produced the clusters with the lowest MCE (<3.2%) in all scenarios considered. Not allowing for the additional heterogeneity within clusters (Yang or Chamroukhi's model) results in a partition with too many clusters and high MCE. Averaging over replicates and dropping the replicate-specific effect (c_{rik}) did not substantially impact clustering results with similar accuracy in prediction of cluster labels between Eq2r1 and Eq2r0. In fact, in a majority of scenarios, averaging over replicates resulted in a slightly lower MCE. Results were similar for data with 50, 125 or 1250 genes per cluster. An increase in between- and within-replicate variability (high error: σ_k^2=0.1 and τ_k^2=0.1) resulted in clusters with slightly higher MCE.

Figure 2.3, summarizes the results of Eq2r1 and Eq2r0 in simulation A compared to K-Means and GMM for all nine underlying assumptions we considered (data with 50 ,125 or 1250 genes per cluster and 4 replicates. Looking at EPEM, for low (σ_k^2=0.01) and high (σ_k^2=0.1) within-replicate variability, an increase in between-replicate error from 0.01 to 0.4 (τ_k^2; rows 1 and 2) did not impact the MCE. However, when the within-replicate variability was very high (0.4), an increase in the

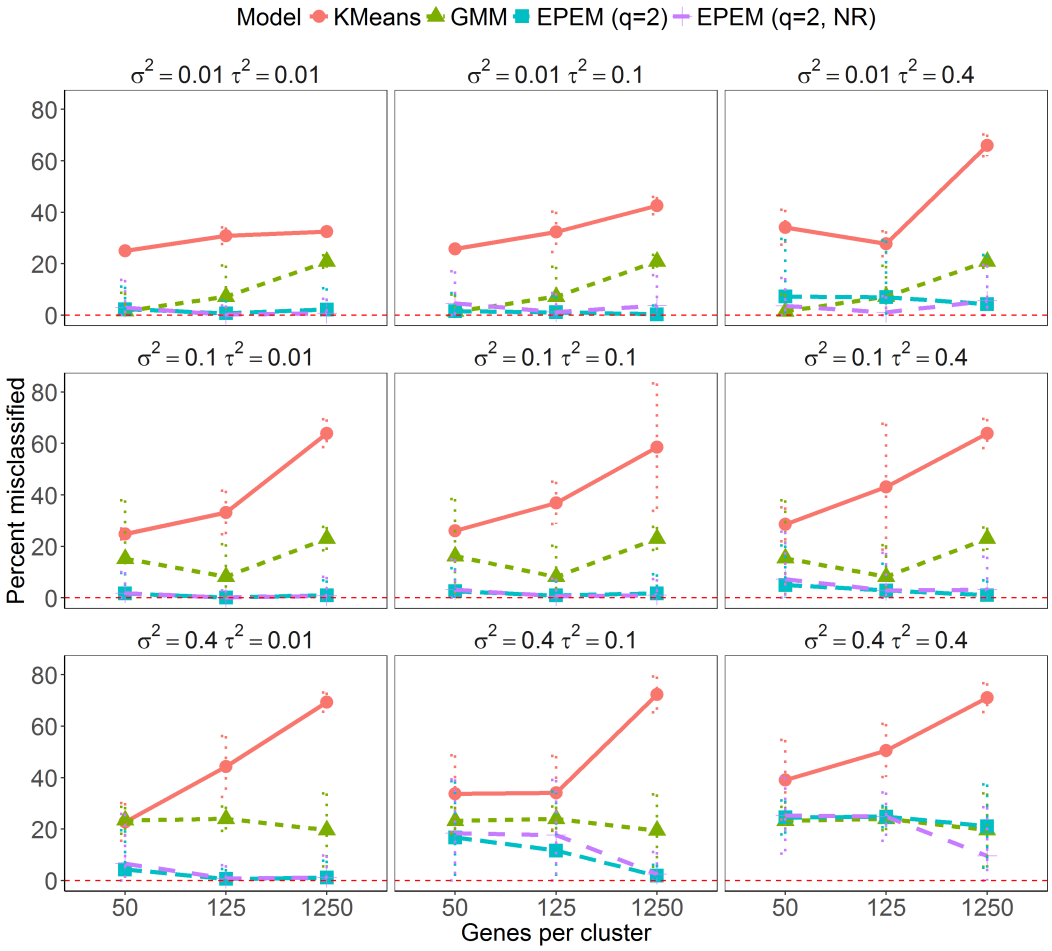

Figure 2.3: Simulation results for K-Means, GMM and EPEM clustering (different underlying models) for data with 4 clusters, 4 replicates and 50, 125 or 1250 genes per cluster over nine different error scenarios. Misclassification error (MCE±SD) over 1000 simulated datasets. (EPEM: Entropy Penalized EM Algorithm; FE: Fixed Effect; RE: Random Effect; GMM: Standard Gaussian Mixture Model; p: FE polynomial order; q: RE polynomial order; NR: drops the replicate-specific RE, c_{rik}, from (2.4).)

between-replicate variability resulted in an increase in MCE of the predicted class labels.

Results were similar when we considered clustering higher order curves (Simulation B), however, clustering data averaged over the replicates (Eq2r0) decreased the MCE by more than 1.7% and performed with the lowest MCE regardless of the number of genes in each. Furthermore, when the data consist of eight clusters, similar results are observed (Supplementary Table B.2).

The estimated parameters for each cluster resulting from Chamroukhi's and Eq2r1 models are given in the Appendix (Table B.3 and B.4) for data with 125 genes/cluster, R=4 and 4 true clusters. In both models, the algorithm did fairly well to estimate the fixed effects parameters (β_k) with low bias. Notice that an increase in predicted cluster number ($\hat{K}$) is proportional to a decrease in the estimated within-replicate variability (σ_k^2). Chamroukhi's model underestimated the within-replicate variability (σ_k^2), resulting in more clusters. The Eq2r1 model was able to accurately estimate α_k, σ_k^2 and τ_k^2. However, for simulation A, the covariance matrix for the gene-specific random effects appear to be biased, potentially attributed to over-specification of q for the linear-curves. In contrast, when the curves were purely cubic, the estimated random-effects are predicted quite well despite the under-specification of q.

2.7.2 Comparison of EPEM to K-Means and Standard Guassian Mixture Model

Table 2.3 also compares KMeans, GMM and EPEM. KMeans performed poorly with MCE > 24% over all scenarios. GMM performed well for smaller datasets (50 or 125 genes/cluster), with an almost perfect clustering in simulation B. However, for higher-dimension data with 1250 genes/cluster, we see that the MCE increases

to more than 14%. Regarding computation time, KMeans had the fastest run-time (<3 minutes) over all scenarios considered. When clustering cubic curves, GMM exhibited the largest increase in run-time. Clustering a dataset with 5000 (1250 genes per cluster; 159 minutes) versus 400 (50 genes per cluster; 1 minute) temporal gene expression profiles was 159 times longer.

In contrast, the EPEM-Eq2r1 was only 65 times longer (353 minutes versus 5 minutes). As N increases, there is evidence that the computation time does not increase as fast for EPEM compared to GMM, most likely attributed to the two-step process when using GMM with an unknown cluster number, which we described previously.

From Figure 2.3, as expected, K-Means performs poorly over all scenarios. GMM performed with only a slightly elevated error compared to the mixed effects models (Eq2r1 or Eq2r0) when the within-replicate variability was in (0.01, 0.10) (Figure 2.3) and cluster size was low (50 or 125 genes per cluster). However, when the within- replicate variability was very high, the MCE quickly increased to more than 20% suggesting that GMM does not work well when the variability is this high. For data with 1250 genes per cluster, GMM performed with higher MCE attributed to an over-estimation of cluster number. Lastly, for high-dimensional data (30 clusters), the EPEM-Eq2r1 resulted in clusters with a lower MCE$\pm$SD compared to GMM (11$\pm$2.6% versus 22$\pm$4.4%, when averaged over 100 simulations).

2.7.3 Varied cluster sizes for EPEM implementation

When the mixing proportions are no longer uniform ($\alpha_k \in$ (0.4,0.3,0.2,0.1)) as opposed to uniform, similar results were observed for EPEM models Eq2r1 and Eq2r0 ((Table 2.4)). When the data consisted of cluster sizes of 80, 60, 40 and 20 genes per

Table 2.4: EPEM clustering results for varied cluster sizes ($\alpha_k \in (0.4, 0.3, 0.2, 0.1)$) (data with 4 clusters, 4 replicates and different underlying error assumptions). Average (SD) predicted cluster number ($\hat{K}$) and misclassification error (MCE %) over 1000 simulated datasets.

		N=200[a], Mean (SD)		N=500[b], Mean (SD)	
Error	Model	$\hat{K}$ (SD)	MCE, %	$\hat{K}$ (SD)	MCE,%
$\sigma^2 = 0.01, \tau^2 = 0.01$	EPEM-Eq0r1 (p=2, q=2)[c]	4.0 (0.29)	1.0 (3.23)	4.0 (0.13)	0.2 (1.26)
	EPEM-Eq2r0 (p=2, q=2, NR)[c]	4.1 (0.36)	1.3 (4.28)	4.0 (0.13)	0.2 (1.39)
	GMM	4.0 (0.16)	0.1 (0.98)	4.7 (0.95)	5.4 (7.27)
$\sigma^2 = 0.01, \tau^2 = 0.1$	EPEM-Eq0r1 (p=2, q=2)	4.1 (0.32)	0.9 (2.99)	4.0 (0.27)	0.6 (2.41)
	EPEM-Eq2r0 (p=2, q=2, NR)	4.2 (0.63)	2.9 (6.31)	4.1 (0.35)	0.8 (3.18)
	GMM	4.0 (0.16)	0.1 (1.14)	4.7 (0.94)	5.4 (7.29)
$\sigma^2 = 0.1, \tau^2 = 0.01$	EPEM-Eq0r1 (p=2, q=2)	4.0 (0.36)	1.4 (3.64)	4.0 (0.21)	0.5 (2.19)
	EPEM-Eq2r0 (p=2, q=2, NR)	4.0 (0.29)	1.0 (3.19)	4.0 (0.18)	0.3 (1.78)
	GMM	4.0 (0.15)	0.3 (1.35)	4.1 (0.33)	1.0 (3.30)
$\sigma^2 = 0.1, \tau^2 = 0.1$	EPEM-Eq0r1 (p=2, q=2)	4.0 (0.38)	1.6 (3.82)	4.0 (0.22)	0.5 (2.11)
	EPEM-Eq2r0 (p=2, q=2, NR)	4.0 (0.40)	1.8 (4.66)	4.0 (0.34)	1.0 (4.45)
	GMM	4.0 (0.15)	0.4 (1.46)	4.1 (0.32)	1.0 (3.23)

EPEM: Entropy Penalized EM Algorithm; FE: Fixed Effect; RE: Random Effect; GMM: Standard Gaussian Mixture Model; p: FE polynomial order; q: RE polynomial order; NR: No replicate-specific RE drops c_{rik} from (2.4).
[a]Resulting in cluster sizes of 80, 60, 40 and 20.
[b]Resulting in cluster sizes of 200, 150, 100 and 50
[c]3% of simulations failed, only 972 iterations reported in this table.

cluster, 3% of the iterations resulted with the covariance matrix of a cluster with very few genes (20 or 40 genes per cluster) to be close to singular, which was also consistent with (Yang et al., 2012). In contrast, for the previous scenario, GMM performs with lower MCE compared to either EPEM models. However, when the cluster sizes increased to 200, 150, 100 or 50 genes per cluster, the two EPEM models perform with higher accuracy than GMM.

2.7.4 Departures from normality for EPEM implementation

When the variabilities are sampled from the Student's T-distribution (Table 2.5), only Eq2r0 and GMM perform with high precision (MCE=0.7%). The other EPEM models we considered are not shown in the table, but performed with similar trends as in Table 2.3. When the errors are simulated from a Gamma distribution, all models perform poorly with MCE >25%.

2.7.5 Effect of replicate number for EPEM implementation

When the effect of replicate number was assessed with the Eq2r1 model in simulation A (Table 2.6), we found that for low-variability data (σ_k^2=0.01, τ_k^2=0.01) an increase in replicate number did not decrease the MCE. In fact, a slight increase was observed. However, for high-variability data (σ_k^2=0.1, τ_k^2=0.1), increasing the replicate number from 2 to 10 lowered the MCE from 8.2% to 2.7% or 1.3% to 0.8% in data with 50 or 125 genes per cluster, respectively.

2.7.6 Convergence of EPEM

The addition of the penalty term no longer ensures the same convergence properties of the EM algorithm. Therefore, to visualize the convergence of the EPEM

Table 2.5: Effect on departures from Normality of error terms on EPEM clustering results (data with 4 clusters, 4 replicates and 125 genes per cluster). Error terms were simulated from either Student's T-distribution (10 df) or Gamma distribution[a]. Average (SD) predicted cluster number ($\hat{K}$) and misclassification error (MCE %) over 1000 simulated datasets.

		Mean (SD)			
Distribution	Model	$\hat{K}$		MCE, %	
Student's T (df=10)	EPEM-Yang[1]	10.1	(1.38)	60.1	(4.54)
	EPEM-Chamroukhi[2]	10.0	(1.23)	59.6	(4.14)
	EPEM-Eq0r1 (p=2, q=0)	3.6	(0.49)	10.7	(12.37)
	EPEM-Eq0r1 (p=2, q=1)	3.0	(0.15)	24.5	(3.76)
	EPEM-Eq0r1 (p=2, q=2)	3.0	(0.23)	25.1	(5.81)
	EPEM-Eq2r0 (p=2, q=2, NR)[3]	4.1	(0.32)	0.7	(2.65)
	GMM	4.1	(0.39)	0.7	(2.68)
Gamma[a]	EPEM-Yang[1]	7.8	(0.98)	46.6	(5.62)
	EPEM-Chamroukhi[2]	7.8	(0.95)	46.5	(5.55)
	EPEM-Eq0r1 (p=2, q=0)	5.7	(0.92)	27.1	(5.23)
	EPEM-Eq0r1 (p=2, q=1)	3.4	(0.73)	25.0	(0.26)
	EPEM-Eq0r1 (p=2, q=2)	3.1	(0.44)	25.0	(0.47)
	EPEM-Eq2r0 (p=2, q=2, NR)[3]	3.1	(0.38)	25.0	(0.87)
	GMM	4.4	(0.93)	25.0	(1.04)

EPEM=Entropy Penalized EM Algorithm; FE: Fixed Effect; RE: Random Effect; GMM: Standard Gaussian Mixture Model; p: FE polynomial order; q: RE polynomial order

[a] $b_{qik} \sim Gamma(3, 10^{q^2})$, $c_{rik} \sim Gamma(3, 5)$, $\epsilon_{trik} \sim Gamma(3, 10)$

[1]Yang's model did not utilize a regression model (no FEs or REs) (Yang et al., 2012).

[2]Chamroukhi's model only included FEs in the regression model (Chamroukhi, 2015).

[3]NR: No replicate-specific RE drops c_{rik} from (2.4).

Table 2.6: Effect of replicate number (R) on the misclassification error using EPEM-Eq2r1 on simulated data with 4 clusters, 50 or 125 genes per cluster, and low or high variability. Mean (SD) over 1000 iterations.

Error[1]	**R**	**50 genes/cluster**, Mean (SD), %	**125 genes/cluster** Mean (SD), %
Low	2	2.39 (4.47)	0.62 (2.39)
	4	2.35 (4.55)	0.86 (2.80)
	10	2.70 (4.78)	0.83 (2.86)
High	2	8.21 (10.49)	1.34 (2.73)
	4	2.56 (4.54)	1.00 (2.96)
	10	2.68 (4.72)	0.83 (2.84)

[1]Low: σ_k^2=0.01, τ_k^2=0.01; High: σ_k^2=0.1, τ_k^2=0.1;

algorithm, we can plot values of the penalized log-likelihood (2.7) and number of clusters for each iteration of the EM-Algorithm. As the number of initial clusters is extremely large, the predicted cluster number for the first few iterations such that $\hat{K} > 50$ are not shown. The red-dashed line corresponds to a cluster number of 4. Figure 2.4 shows these results for one simulated dataset with high variability, for different underlying models (Yang, Chamroukhi or Eq2r1). For all models, the penalized-log likelihood becomes flat after 125, 130 or 140 iterations for Yang, Chamroukhi or Eq2rq models, respectively. The Yang or Chamroukhi model converges in fewer iterations compared to Eq2r1, but both resulted in a higher MCE (>37%) and larger predicted cluster number as opposed to Eq2r1 with no genes being misclassified. We can see that the number of clusters at each subsequent iteration of the EM drops quickly.

2.7.7 Application of EPEM to fracture healing study

From Figure 2.5, 22 clusters were predicted from the clustering algorithm. Clusters 1-10 showed an initial increasing trend. Clusters 11-13 showed relatively flat clus-

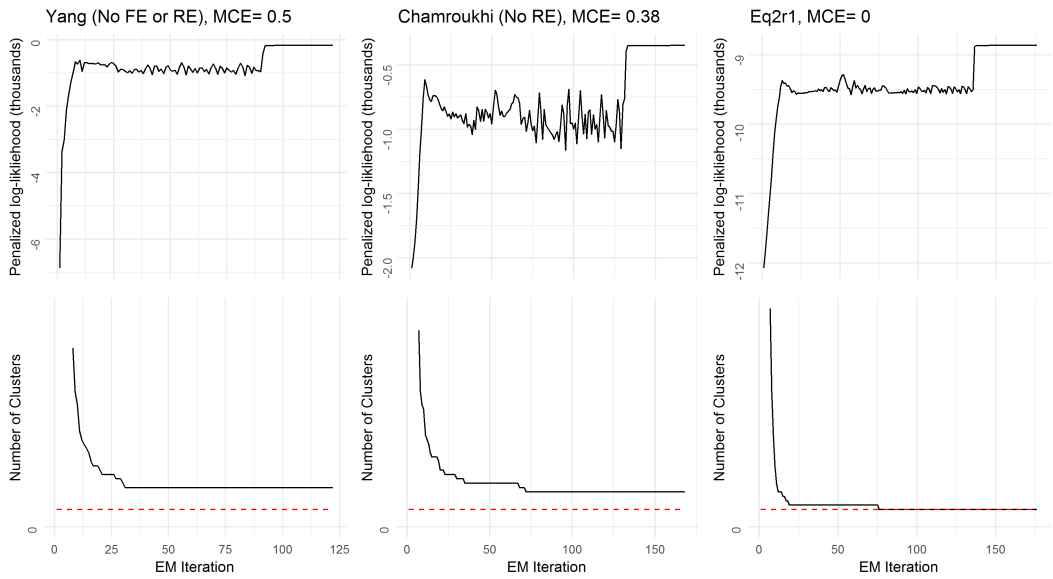

Figure 2.4: Convergence of objective function and the number of clusters (when predicted cluster number is less than 50) for models Yang (no fixed-effects (FE) or random-effects (RE)), Chamroukhi (no RE) and Eq2r1 for iteration 1 of the EPEM algorithm with 4 clusters, 125 genes per cluster and data with high variability (σ_k^2=0.1, τ_k^2=0.1).

ters as evidenced by the small predicted β coefficients (Table B.5). Clusters 14-22 represent clusters with an initial decreasing trend. Within-replicate variability ($\hat{\sigma}_k^2$) for each cluster ranged from 0.16 to 0.46. Between-replicate variability ($\hat{\tau}_k^2$) ranged from 0.02 to 0.66. Recall that the between-replicate variability represents the variability between strains (AJ, B6 and C3H), suggesting a sizable amount of variation in gene-expression between the three strains. Total run time took about 3 days (71 hours).

2.8 DISCUSSION

In this chapter, we proposed an entropy penalized EM algorithm with mixtures of ME regression models to cluster data with additional variability (from biological replicates and repeated measurements over time). We found that the addition of

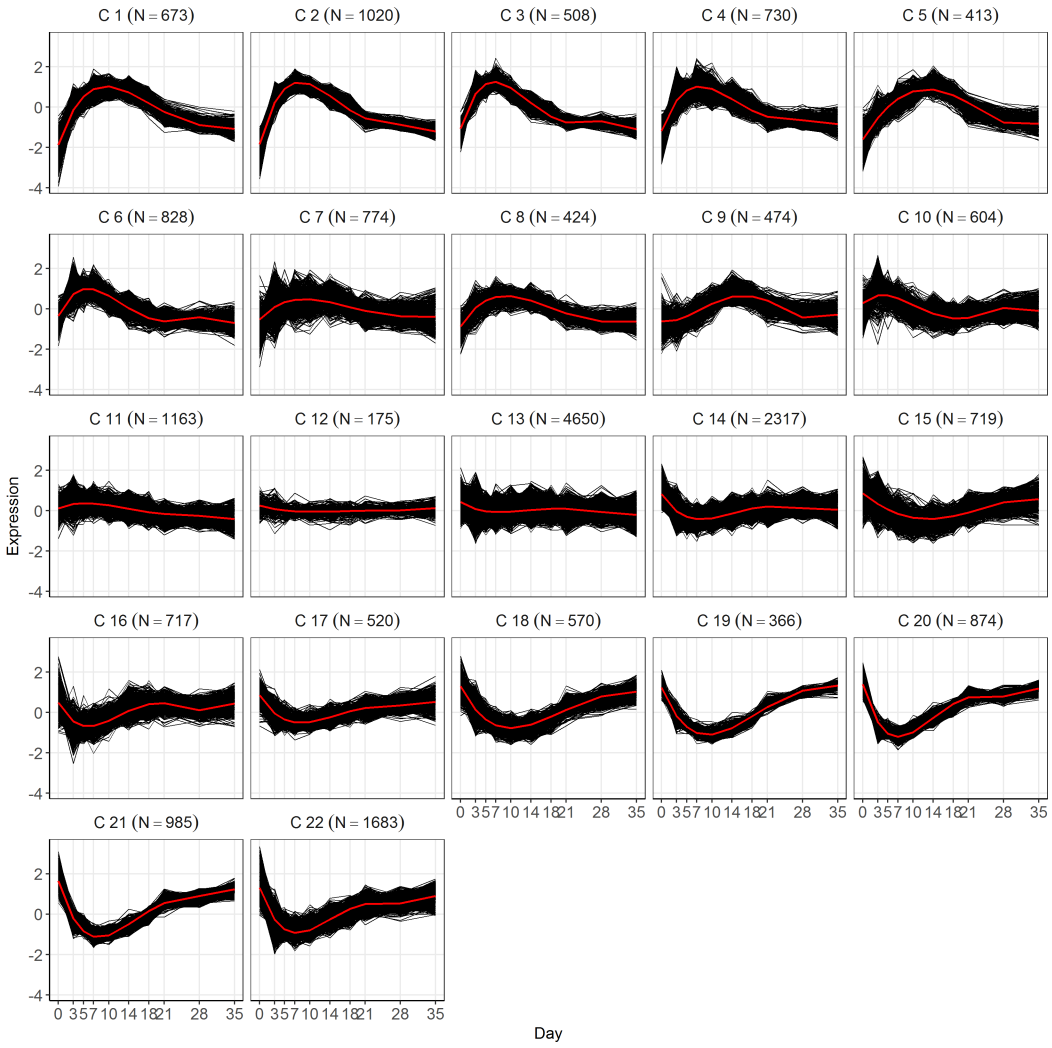

Figure 2.5: Clustering results from fracture healing microarray study with entropy penalized EM-algorithm Eq2r1 (p=4, q=2, strain (replicate)-specific RE). Each plot represents temporal gene-expression profiles clustered into the same group. Total run time was 71 hours. (C: Cluster)

REs in our mixture model decreased the misclassification error by the clustering algorithm compared to mixtures of FE models (Chamroukhi, 2015) and other more popular methods (Yang et al., 2012) (Fraley et al., 2012) (Hartigan & Wong, 1979).

In data with high within- and between-replicate error, we found that having more replicates slightly lowered the MCE. Increasing the cluster size drastically decreased the MCE, but similar trends were observed. Furthermore, in all scenarios, no big differences in the accuracy of the predicted class labels were observed when we ignored the replicate-specific RE, c_{rik} (Eq2r1 versus Eq2r0), suggesting that averaging over replicates is an effective way to reduce the dimension of the data and prevent over-specifying the model without a sacrifice in accuracy.

When we assessed data with different cluster sizes, we found that clusters with very few number of genes could result in singularity issues in the covariance matrix estimation of the EPEM algorithm, which is consistent to findings from Yang (Yang et al., 2012). Yang's solution was to use a constrained covariance matrix to circumvent this, which is an option we can implement in the future. However for the purposes of this manuscript with a primary application to gene expression clustering with thousands of genes, we believed that having large enough cluster sizes was a reasonable assumption to make.

In general, GMM performed well, however situations with a very high variability (i.e. σ_k=0.4), sample size (i.e. 1250 genes per cluster) or cluster number (i.e. K=30) resulted in classifications with reduced accuracy. When the data consist of errors with a heavy tail, suggestive of the possibility of outliers, GMM and Eq2r0 both perform equally well. Averaging over the replicates was a way to reduce the effect of the outliers. Models accounting for the replicates (i.e. Eq2r1) performed poorly with high MCE. When the data are heavily skewed (Gamma), none of the

methods performed well. In this case, an extension to a mixture model defined by gamma distributions would most likely result in better clusters (Lakshmi & Vaidyanathan, 2016). The performance of any model-based clustering method is dependent on how well the model fits to the data.

For large datasets, despite the high accuracy of predicted class labels, a major limitation of this method is that the computation time is quadratic in growth for datasets with a large number of genes. Clustering of the fracture healing data took about three days to run. With the rise of big-data and availability of high-dimensional data, the use of a clustering algorithm that can handle different sources of variability would be important to be able to determine clustering patterns that may not be easily obtained otherwise. We have shown that using the methods proposed in this paper, we can dramatically improve the accuracy of predicted class labels for a large set of genes. In the next chapter, modifications to the EPEM are considered that aim to substantially decrease the computational cost for these big datasets.

CHAPTER 3

Model selection and considerations for high-dimensional data

3.1 BACKGROUND

Model-based clustering using mixture models is a popular tool to obtain homogeneous groups within temporal data. The flexibility and probabilistic framework of model-based clustering are a few benefits that are associated with using these techniques. With an increase in technological advances, the availability of high-dimensional data is becoming more prevalent. Unfortunately, the benefits of using model-based clustering methods are now being outweighed by its computational cost. To counteract this, in many gene expression analyses, a subset of the genes is used in the cluster analysis in an attempt to decrease the size of the data to cluster (Ng et al., 2006) (Celeux et al., 2005). However, the use of different thresholds of inclusion may affect the results leading to clusters derived from incomplete data.

In the previous chapter, to cluster temporal gene-expression data, a mixed-effects (ME) model based cluster algorithm was proposed using an entropy penalized EM algorithm (EPEM). The additional penalty term created a data-driven algorithm to simultaneously estimate the number of clusters and the cluster labels for a set of temporal gene-expression profiles. Multiple runs of the clustering algorithm over a range of cluster numbers is no longer necessary, which many traditional algorithms utilize (i.e. K-Means and the standard Gaussian mixture model (GMM)). From several simulation studies, the addition of random effects (RE) into the regression model drastically decreased the misclassification error (MCE) of our predicted cluster labels.

However, for high-dimensional data, the computational cost was high. Datasets

with more than 10,000 temporal gene expression profiles took many hours to run despite the low MCE. The long run-time was attributed to the initialization step of the algorithm, which initialized with the maximum number of possible clusters (a dataset with 10,000 genes would initialize with 10,000 clusters). As a result, estimation in the first few iterations of the algorithm is very expensive. To counteract this cost, two methods were investigated to reduce the up-front cost of the EPEM, a split-clustering algorithm (S-EPEM) and a modified-initialization algorithm (I-EPEM). S-EPEM clusters the data in groups defined by genes with the same predicted p-th order polynomial function. However, the final set of cluster labels is highly dependent on how the groups are defined. I-EPEM uses a pre-specified naive grouping based on estimated p-th order polynomial coefficients in an effort to decrease the initial number of clusters.

Furthermore, when working with polynomial regression models, we must acknowledge that these models may be too simple to fully capture highly non-linear time trends. However, we believe that they can be a good approximation to temporal gene-expression time profiles used for the purpose of exploratory cluster analysis. It follows that an important step in polynomial regression models is to choose the correct order to best represent the temporal pattern for each gene. In the previous chapter, we had assumed that the fixed-effect order, p, of our polynomial model to be known, which is usually not the case in real data. Therefore, in this chapter, a simulation study was conducted to compare different model selection methods to determine the optimal p using AIC, BIC, leave-one-out CV (LOOCV) and repeated K-fold CV with 10 folds and 100 repeats (Burnham & Anderson, 2003) (Shao, 1993). Additionally, a dimension reduction tool, singular value decomposition (SVD), is also used in comparison to determine the optimal p to be

used in the EPEM algorithm (Storey et al., 2005) (Wall et al., 2003).

3.2 METHODS

3.2.1 The Data

Assume that the data consist of N temporal gene-expression profiles with measurements obtained at T time points with R replicates at each time-point. Therefore, each temporal gene-expression profile has T$\times$R measurements. The goal is to cluster the N temporal gene-expression profiles into K groups using a modified version of the entropy penalized EM algorithm with Gaussian mixtures of polynomial mixed effects models. The modified version would ideally perform with a faster computation time without a sacrifice in accuracy.

3.2.2 Model Selection to optimize FE polynomial order

From our simulation studies in Chapter 2, the EPEM algorithm set the FE order, p, from (2.4) to be the maximum order, $p^{(max)}$, used to simulate each curve (in simulation A and B, $p^{(max)}$=2 or 3, respectively; Table 2.2). However, in the majority of scenarios, the dimensionality or order of the data is unknown. In order to determine the order of our data, a comparison between model-selection techniques was conducted to determine the predicted order number ($\hat{p}$) using AIC, BIC and cross-validation (CV) methods.

Two model selection scenarios were considered. First, model selection was performed using all TR observed observations from gene i. Model selection methods using AIC, BIC and cross-validation were considered and compared. The cross validation methods we considered were leave-one-out cross validation
(LOOCV) and repeated K-fold cross-validation (K-fold CV). LOOCV trains the

data on TR-1 observations and calculates the squared-error of the left-out observation. Repeated ten-fold CV was performed with 10 folds and 100 repeats, where the 10 folds are determined by randomly dividing the data into 10 folds across all TR observations and uses data from 9 folds as the training sample and calculates the squared errors for observations from the left out fold (test sample). It does this for all folds and is repeated 100 times to obtain the average squared-error.

Second, instead of using the observed gene-expression profiles, the model selection will be performed on the top l eigenvectors ($l = 1, \ldots, L$ where L $\leq$ TR) obtained by decomposing the standardized gene-expression data matrix using singular value decomposition (SVD). SVD was used because it has been shown to be effective in reducing the dimension of a gene-expression data matrix to extract the top subspaces that explain the most variability of the data (Storey et al., 2005) (Wall et al., 2003). Model selection techniques can then be used to estimate the order of each of these subspaces to find $p^{(max)}$ that is sufficiently large enough to explain the patterns seen in the original data-matrix. See **Appendix A.4** for more details on SVD.

The steps for model selection are as follows:

1. For each gene or eigenvector, fit fixed-effects polynomial regression models with p $\in$(1, 2, 3, 4).

2. Determine AIC, BIC or MSE ($\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$) from CV for each p.

3. Select model with lowest AIC, BIC or MSE to determine $\hat{p}_i$ for each gene.

4. Repeat steps 1-3 for all genes or eigenvectors.

5. Determine maximum p over all genes (i=1,..., N; $\hat{p}^{(max)} = max_i \hat{p}_i$) or eigenvectors (l=1,...,L; $\hat{p}^{(max)} = max_l \hat{p}_l$).

$\hat{p}^{(max)}$ is the optimal FE polynomial order to be used in the EPEM clustering algorithm.

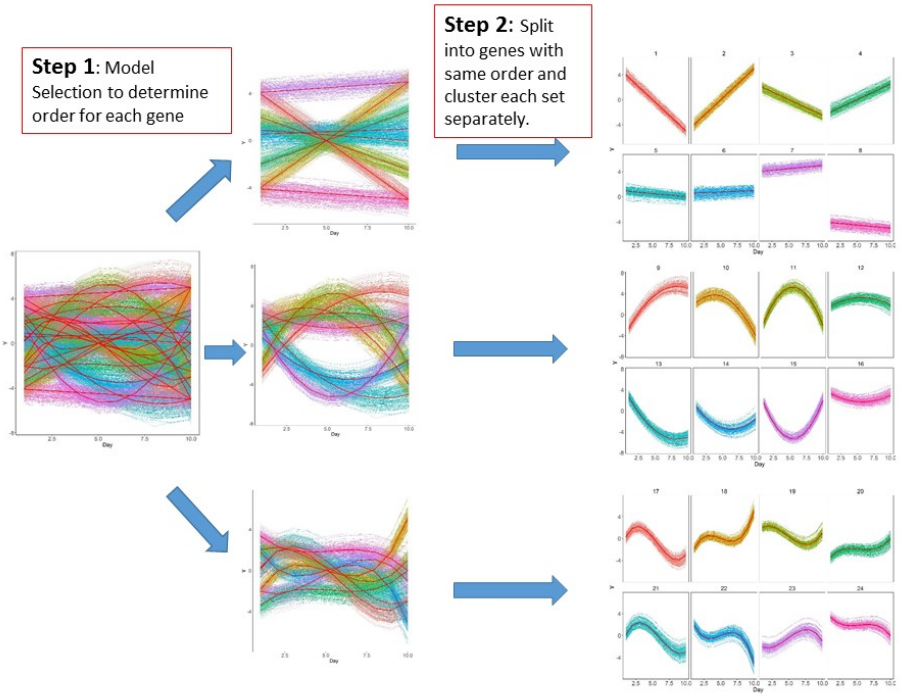

Figure 3.1: Split-clustering EPEM schematic

3.2.3 Split-clustering

Intuitively, setting p on a gene-by-gene basis should result in accurate results to prevent under- or over-specification of our mixed effects model. However, to accomplish this in the clustering framework, we would need to split the data into subsets where each subset includes genes that can be represented by the same order polynomial function. Separate runs of EPEM with models with varying fixed effect order (p) can then be used to cluster the data. For example, if you split your data into three sets of curves (linear, quadratic or cubic), the clustering algorithm would be run for p=1 for the linear set, p=2 for the quadratic set, and p=3 for the

cubic set. Using model selection methods (i.e. AIC, BIC or CV), the predicted order for gene i, $\hat{p}_i$, can be determined for each gene. The data can now be subdivided into sets with the same $\hat{p}_i$. The clustering algorithm for each set will be run with the corresponding predicted order, $\hat{p}_i$. Simulations where subsets were defined using the true polynomial order was also used. Results performed similarly to EPEM, but were not reported here. The MCE from the Split-EPEM algorithm (S-EPEM) will be compared to the MCE from the original EPEM algorithm proposed in **Chapter 2**. A schematic of the procedure is given in 3.1.

3.2.4 Modified Initialization

An alternative to a split-clustering technique is to modify the initialization of the algorithm to start with fewer clusters (as opposed to N). If a p-th order fixed-effects polynomial regression model is fitted to each of the N gene expression time profiles (each with TR observations), a set of estimated beta-coefficients can be obtained. The genes can be grouped based on the set of estimated coefficients. Using the p+1 regression coefficients ($\hat{\beta}_{0i}$, $\hat{\beta}_{1i}$, ..., $\hat{\beta}_{pi}$), we hierarchically grouped the N genes with the following steps (Figure 3.2):

1. Run a p-th order polynomial regression for each of the N temporal gene expression profiles.

2. Extract the β coefficients ($\hat{\beta}_{0i}$, $\hat{\beta}_{1i}$, ..., $\hat{\beta}_{pi}$)

3. Using $\hat{\beta}_{pi}$, split the genes into five groups based on their percentiles (quintiles).

4. Within each of the five groups, split the genes into 5 groups based on their $\hat{\beta}_{(p-1)i}$.

5. Repeat 4 for $\hat{\beta}_{(p-2)i}, \ldots, \hat{\beta}_{0i}$.

We chose to use quintiles resulting in 5^{p+1} clusters instead of N. Tertiles were also considered (3^{p+1} initial clusters), but the reduction in initial clusters resulted in results with a much higher MCE (11 splits were also considered, but there was no big difference in reduction of runtime). Note that the dataset must be of sufficient size such that $5^{p+1} <$ N.

3.2.5 Simulation Study

1000 datasets were generated with 24 clusters, 4 replicates per gene, 125 genes per cluster, and varied within- (σ_k^2) and between- (τ_k^2) replicate variability (low: $\sigma_k^2 = 0.01, \tau_k^2 = 0.01$; high: $\sigma_k^2 = 0.1, \tau_k^2 = 0.1$) from a fully-specified mixed effects model defined in (3.1). Figure 3.3 represents one simulated data of the temporal gene

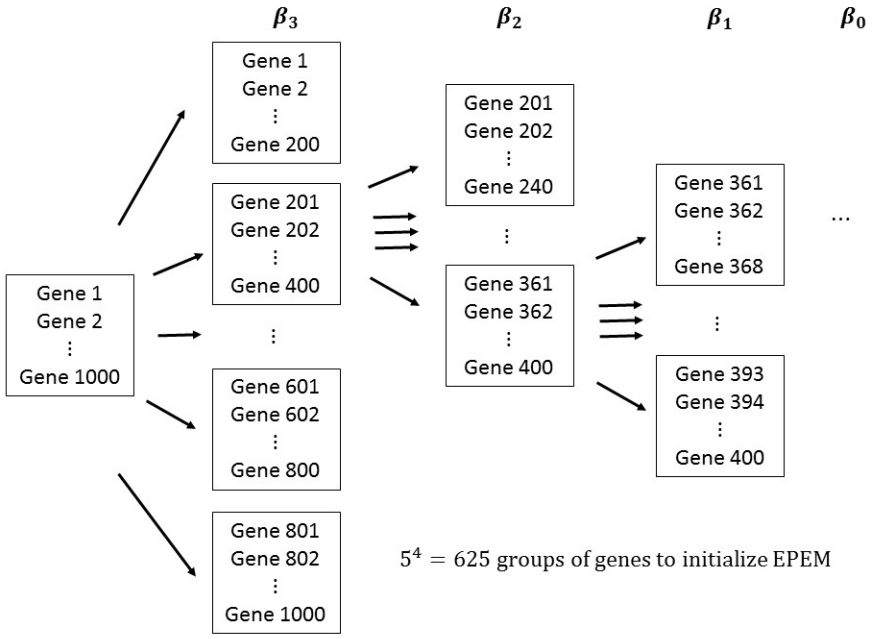

Figure 3.2: Modified initialization EPEM schematic with 5 splits at each level

expression curves, each with 8 clusters exhibiting linear (p=1), quadratic (p=2), and cubic (p=3) trends. The table of parameters used to simulate the data are given in 3.1.

$$Eq2r1 : y_{trik} = \beta_{0k}+\beta_{1k}x_{tri}+...+\beta_{pk}x_{tri}^{p}+b_{0ik}+b_{1ik}x_{tri}+\ldots+b_{qik}x_{tri}^{q}+c_{rik}+\epsilon_{trik} \quad (3.1)$$

$$Eq2r0 : y_{tik} = \beta_{0k} + \beta_{1k}x_{ti} + ... + \beta_{pk}x_{ti}^{p} + b_{0ik} + b_{1ik}x_{ti} + \ldots + b_{qik}x_{ti}^{q} + \epsilon_{tik} \quad (3.2)$$

To assess each model selection criteria, the percent of correct predictions of p for AIC, BIC and CV is determined for each dataset. The average and standard deviation (SD) over all 1000 datasets is reported. EPEM was run with p=2, 3 or 4 and q=2 to assess the effect of under-specifying (p=2) versus over-specifying (p=4) the mixed effects model on the clustering accuracy. The performance of S-EPEM was assessed by splitting the data into three groups based on the true order (p = 1, 2 or 3) or the predicted order ($\hat{p}_i$=1, $\hat{p}_i$=2, or $\hat{p}_i \in \{3, 4\}$; Only a few genes had $\hat{p}_i$=4). EPEM was run separately for the three subsets of data with p=1, 2 or 3, respectively. The performance of I-EPEM was assessed using a mixed effects (ME) regression model with p=3 and q=2 by pre-grouping the data into 625 initial clusters (as opposed to 3000). EPEM, S-EPEM and I-EPEM are run with models with and without a replicate-specific random effect (c_{rik}; Eq2r1 (3.1) or Eq2r0 (3.2)). If you recall, Eq2r0 requires the clustering to be done on data that is averaged over the replicates. The average and SD of the predicted cluster number ($\hat{K}$), accuracy (MCE, %), and run-time (hours) over the 1000 datasets are reported.

Convergence of the algorithm was assessed by the FE coefficients such that $max_k||\beta_k^{(s)} - \beta_k^{(s+1)}|| < \epsilon = 10^{-3}$. We had initially used 10^{-4}, but due to the large number of clusters, convergence time was very long; However, for 100 iterations,

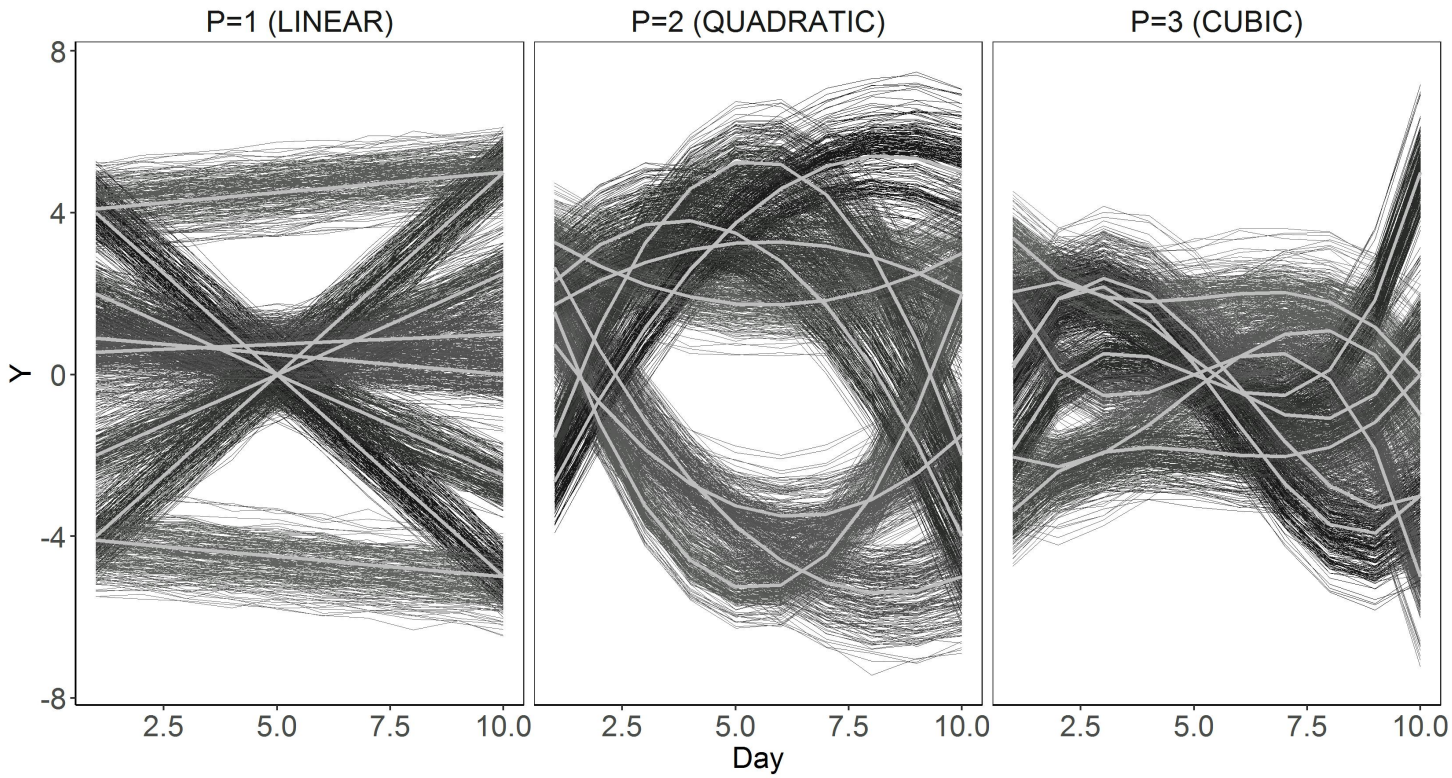

Figure 3.3: Simulated dataset with 24 clusters, 125 temporal gene expression curves per cluster and 4 replicates for each gene with low ($\sigma_k^2 = 0.01, \tau_k^2 = 0.01$) error. The 24 clusters consisted of linear, quadratic and cubic curves, each with eight clusters

Table 3.1: Simulation parameters to obtain datasets used in simulation studies with equal mixing proportions (α_k=0.25) for each cluster. Datasets were generated with 125 genes per cluster, 4 replicates and low ($\sigma_k^2 = 0.01, \tau_k^2 = 0.01$) or high ($\sigma_k^2 = 0.1, \tau_k^2 = 0.1$) within- or between-replicate variability and gene-specific random effect correlations specified by $\rho(b_0, b_1) = -0.5$, $\rho(b_0, b_2) = 0.4$, $\rho(b_1, b_2) = -0.8$, $\rho(b_0, b_3) = 0.1$, $\rho(b_1, b_3) = -0.2$, and $\rho(b_2, b_3) = 0.6$.

Cluster	β_0	β_1	β_2	β_3
1	5	-1	–	0
2	-5	1	–	0
3	2.5	-0.5	–	0
4	-2.5	0.5	0	0
5	1	-0.1	0	0
6	0.5	0.05	0	0
7	4	0.1	0	0
8	-4	-0.1	0	0
9	-5	2.5	-0.15	0
10	1	1.5	-0.2	0
11	-5	3.8	-0.35	0
12	1	0.8	-0.07	0
13	5	-2.5	0.15	0
14	2.5	-1.9	0.15	0
15	5	-3.8	0.35	0
16	4	-0.8	0.07	0
17	-3	4.2	-1	0.058
18	-5	4	-0.9	0.06
19	1	1.5	-0.5	0.035
20	-5	2	-0.4	0.025
21	-3	4	-0.88	0.048
22	5	-4	0.9	-0.06
23	-1	-1.5	0.5	-0.035
24	5	-2	0.4	-0.025

no impact on results was seen when $\epsilon = 10^{-3}$.

Lastly, 1000 datasets from Simulation A with 4 clusters and 4 replicates was used (Table 2.2) to compare the effect of sample size (50, 125 or 1250 genes/cluster) and underlying error assumption (nine pairwise combinations of $\{0.01, 0.1, 0.4\}$) on the clustering results from EPEM and I-EPEM (p=2, q=2 with and without replicate-specific RE; initializing with 125 clusters compared to 200, 500 or 5000). S-EPEM was not assessed in this way as the results are expected to perform similarly to EPEM for all scenarios as the algorithm itself is not altered.

3.2.6 Application to the fracture healing study

Similar to the previous chapter, I-EPEM algorithm with an underlying Eq2r1 underlying model (p=4 determined by SVD; q=2; (3.3)) was used to cluster the fracture healing data. For each strain and time-point, the data were averaged over the 3 replicates. The replicate-specific random effect, c_{rik}, was included in the model to account for strain-specific variability. The expression values were standardized by the gene-specific mean and standard deviation over all strains. Additional information on the fracture-healing study are given in **Appendix A.1**. The results and run-time are compared between I-EPEM and EPEM.

$$y_{trik} = \beta_{0k}+\beta_{1k}x_{tri}+\beta_{2k}x_{tri}^2+\beta_{3k}x_{tri}^3+\beta_{4k}x_{tri}^4+b_{0ik}+b_{1ik}x_{tri}+b_{2ik}x_{tri}^2+c_{rik}+\epsilon_{trik} \quad (3.3)$$

Additionally, S-EPEM is run by separating the genes into sets of linear (p=1), quadratic (p=2), cubic (p=3) and quartic curves (p=4) using model selection from **Section 3.2.2** and BIC. Similarity of cluster labels between the two algorithms are compared using the adjusted rand index (ARI; see Appendix A.6 Equation (A.18)) (Hubert & Arabie, 1985) and a confusion matrix of cluster labels. The ARI is an

adjusted measure of agreement between two sets of cluster labels. It determines the proportion of pairs of objects that are in the same cluster in both sets of labels, and adjusted to have an expected value of 0. When two sets of labels perfectly agree, the ARI is 1.

3.3 RESULTS

3.3.1 Model selection to optimize fixed effect polynomial order

Observed gene-expression measurements

For each simulation, $\hat{p}_i$ was determined for each gene expression profile considered from each dataset across 1000 simulated datasets. This resulted in 10^6 linear, cubic and quadratic curves where model selection is performed for a total of 3×10^6 curves. The percent of predictions where p = 1, 2, 3 or 4 for linear, quadratic and cubic curves for each criterion (AIC, BIC, LOOCV and 10-fold CV) was determined and reported in Table 3.2.

BIC performed with the highest accuracy with 97% correct predictions, which could be attributed to the penalization of higher order models (Friedman et al., 2001). For AIC and CV methods performed with about 90% accuracy for each type of curve. In fact, it has been shown that AIC and LOOCV are asymptotically equivalent (Stone, 1977). AIC and CV methods tended to choose more complex models (10% were over-fit), compared to BIC, which penalized model complexity more heavily (3% were over-fit). Note that BIC = log(number of observations)(number of parameters)-2*Log-Likelihood. AIC = 2(number of parameters)-2*Log-Likelihood. Therefore, when the number of observations is large enough (>7), the penalty for BIC will be much larger than AIC, therefore reducing the tendency to over-fit the model. However, regardless of the criterion used, the optimal p, $\hat{p}^{(max)}$, from all

Table 3.2: Percent of times model selection criteria using LOOCV, 10-fold CV, AIC or BIC chose a particular polynomial order to best fit the observed temporal gene-expression curve. Results are presented by type of observed curve originating from 10^6 linear, 10^6 quadratic, and 10^6 cubic temporal gene expression curves (data with low error: $\sigma^2 = 0.01, \tau^2 = 0.01$).

Type of Curve	**($\hat{p}$)**	**LOOCV**	**10-Fold CV**	**AIC**	**BIC**
Linear (p_{true}=1)	4	1.4%	1.5%	2.0%	0.1%
	3	2.7%	2.9%	3.1%	0.5%
	2	6.9%	6.9%	7.3%	2.2%
	1	89.0%	89.0%	87.6%	97.2%
Quadratic (p_{true}=2)	4	2.8%	2.8%	3.8%	0.5%
	3	6.9%	6.9%	7.8%	2.3%
	2	90.2%	90.3%	88.4%	97.2%
Cubic (p_{true}=3)	4	7.5%	7.5%	9.1%	2.5%
	3	92.5%	92.5%	90.9%	97.5%

p: fixed-effect polynomial order; p_{true}: true polynomial order used to simulate the data; $\hat{p}$: predicted polynomial order from model selection.

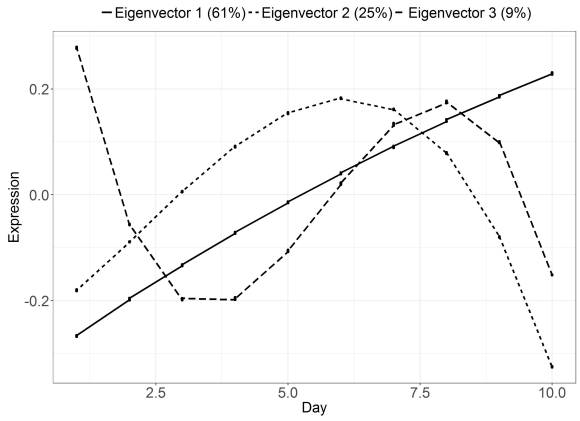

Figure 3.4: Top three eigenvectors obtained from singular value decomposition of the gene-expression data matrix from iteration 7. Together, the three eigenvectors explained $> 90\%$ of the variability of the data.

the genes within each simulation was 4. In other words, for each simulation, at least one gene resulted in $\hat{p}_i$=4.

Eigenvectors from Singular Value Decomposition

Using SVD, we chose L to be three as the top 3 eigenvectors explained more than 90% of the variability in the data. When the three eigenvectors were plotted over time (Figure 3.4), three distinct patterns emerge (linear, quadratic and cubic functions).

To further explore the relationship of the eigenvectors with the individual temporal gene-expression profiles, Pearson correlations between each of the three eigenvectors and gene-expression profiles were computed for one dataset (Figure 3.5) and presented in a 3-dimensional plot. Each point in the 3-dimensional space represents the correlation between one temporal gene-expression profile with eigenvector 1, 2 and 3. Groups of genes that aggregate at the perimeter of each axis (around -1 and 1) were highly correlated with the corresponding eigenvector. Color

coding the genes based on their temporal order of expression (i.e. linear, quadratic or cubic) reveals expected patterns between the eigenvectors and the genes. Genes with linear curves (blue) were strongly correlated eigenvector 1 only. Genes with quadratic curves (pink) were strongly correlated with eigenvector 2, with some genes showing correlation with eigenvector 1. Genes with cubic curves (black) were correlated with all three eigenvectors with varying degrees. Similar results were seen for all 1000 simulations. The results suggested the ability of SVD to extract subspaces from the observed data matrix that can sufficiently represent the overall patterns seen in this data. Performing model selection using BIC on the three eigenvectors for each simulation resulted in $\hat{p}^{(max)}$=3 was chosen 96% of the time over the 1000 simulations.

EPEM implementation

If we performed model selection on each observed gene-expression profile, where $\hat{p}^{(max)}$=4, the EPEM algorithm would be run with p=4. If model selection was performed on the top 3 eigenvectors, the EPEM algorithm would most likely be run with p=3. The effect of FE order specification, p, on the clustering results is shown in Table 3.3 for a model specified by Eq2r1 (3.1) or Eq2r0 (3.2). It can be seen that an over-specification of the model with p=4 did not negatively impact the accuracy of the predicted cluster labels for either model regardless of variability. However, if p is underspecified with p=2, the accuracy of the clustering algorithm is low with >23% of the temporal gene profiles being misclassified.

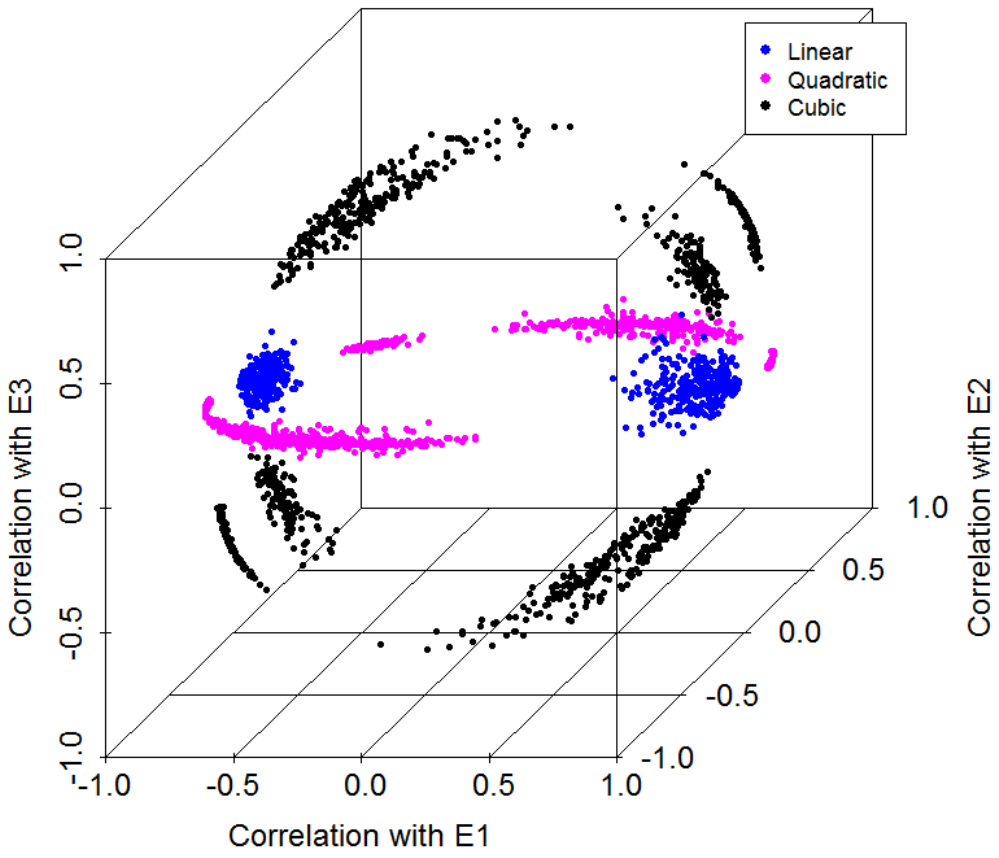

Figure 3.5: Correlation of the top 3 eigenvectors with each observed temporal gene-expression profile from one simulation.

Table 3.3: Simulation Results with entropy penalized EM algorithm (EPEM), split-clustering EPEM (S-EPEM), and modified-initialization EPEM (I-EPEM) (data with 24 clusters, 4 replicates and 125 genes per cluster) for models with (Eq2r1) and without (Eq2r0) a replicate-specific random-effect (RE; c_{rik}). Average (SD) predicted cluster number ($\hat{K}$), misclassification error (MCE %) and convergence time (hours) over 1000 simulated datasets.

A. Model Eq2r1 (including a replicate-specific RE)

			Mean (SD)		
Error	Algorithm	Model	$\hat{K}$	MCE, %	RunTime, hours
Low	EPEM	p=2, q=2	17.1 (1.99)	29.2 (7.85)	3.06 (2.44)
		p=3, q=2	23.5 (1.40)	4.1 (2.98)	2.31 (1.29)
		p=4, q=2	23.6 (1.20)	4.1 (2.91)	2.62 (1.43)
	I-EPEM	p=3, q=2	24.1 (1.48)	2.8 (2.58)	1.05 (0.65)
	S-EPEM	Predicted p and q	25.2 (1.52)	4.6 (2.23)	0.35 (0.27)
High	EPEM	p=2, q=2	18.9 (1.41)	22.6 (4.95)	5.29 (2.64)
		p=3, q=2	23.7 (1.40)	3.7 (2.66)	4.56 (1.93)
		p=4, q=2	23.8 (1.36)	4.0 (2.79)	5.04 (2.14)
	I-EPEM	p=3, q=2	24.1 (1.21)	2.5 (2.51)	2.05 (0.91)
	S-EPEM	Predicted p and q	25.5 (2.00)	5.3 (2.39)	0.68 (0.22)

B. Model Eq2r0 (no replicate-specific RE)

			Mean (SD)		
Error	Algorithm	Model	$\hat{K}$	MCE, %	RunTime, hours
Low	EPEM	p=2, q=2	18.2 (1.47)	24.9 (5.73)	1.58 (0.92)
		p=3, q=2	23.7 (1.04)	3.2 (2.55)	1.05 (0.50)
		p=4, q=2	23.7 (1.03)	3.1 (2.61)	1.01 (0.55)
	I-EPEM	p=3, q=2	24.0 (1.40)	3.2 (2.69)	0.66 (0.60)
	S-EPEM	Predicted p and q	24.7 (1.07)	3.9 (2.12)	0.17 (0.07)
High	EPEM	p=2, q=2	18.8 (1.29)	22.5 (4.81)	2.90 (1.32)
		p=3, q=2	23.3 (0.95)	4.0 (3.20)	3.76 (1.73)
		p=4, q=2	23.3 (0.92)	3.8 (3.05)	3.61 (1.72)
	I-EPEM	p=3, q=2	23.9 (1.08)	2.6 (2.48)	1.92 (1.10)
	S-EPEM	Predicted p and q	24.5 (1.21)	4.5 (2.27)	0.58 (0.16)

EPEM: 3000 clusters used for initialization
I-EPEM: Pre-grouped into 625 clusters for initialization
S-EPEM: Split into 3 sets to run with p=1, q=1 or p=2, q=2 or p=3, q=2

3.3.2 Split-clustering

Including a replicate-specific RE and using S-EPEM, the MCE was slightly higher than EPEM (low error: 4.6% versus 4.1%; high error: 3.7% versus 5.3%; Table 3.3A)), and similarly when we ignored the replicate-specific RE (Table 3.3B). As expected, when another layer of classification is performed to split the data into subsets, a slight decrease in accuracy of predicted cluster labels was observed. However, the benefit of this algorithm is in the run-time. The runtime for S-EPEM was only 0.35 hours (21 minutes) versus >2 hours for EPEM (Table 3.3A).

3.3.3 Modified Initialization

Using the modified-initialization algorithm (I-EPEM) for data with 24 clusters, we found that a decrease in runtime was not as dramatic as what we saw for S-EPEM, but in the majority of instances, we were able to decrease run-time by about half (Table 3.3). In contrast to S-EPEM, the I-EPEM algorithm performed with high accuracy over all scenarios considered (MCE <3.2%).

As I-EPEM performed with similar or lower accuracy than EPEM, but the algorithm is now altered due to the initialization piece, a further comparison was conducted to determine if there were certain scenarios where I-EPEM did not perform well (Table 3.4). When sample size was small (50 or 125 genes per cluster), I-EPEM performed similarly to EPEM regardless of underlying error or mixed effects model used (with or without replicate-specific RE) with no beneficial differences in runtime as EPEM works just as quickly (<10 minutes). When the data consisted of 1250 genes per cluster, I-EPEM runs must faster (<18 minutes) compared to EPEM, which took more than 150 minutes (2.5 hours) to run.

Using a ME model with a replicate-specific RE ((3.1) with p=2, q=2; Table 3.3A),

Table 3.4: Misclassification error (MCE(SD)% over 1000 iterations) from clustering data (4 clusters, 4 replicates and 50, 125 or 1250 genes per cluster) with 2 mixed effects models (Eq2r1: p=2, q=2, replicate-specific RE; Eq2r0: p=2, q=2, no replicate-specific RE) and clustering algorithms (EPEM or I-EPEM) for different error scenarios. (EPEM: Entropy Penalized EM Algorithm; I-EPEM: Modified Initialization EPEM; Rep: Replicate; RE: Random Effect)

A. Including a replicate-specific RE (Eq2r1)

		50 genes/cluster, Mean (SD)			**125 genes/cluster**, Mean (SD)			**1250 genes/cluster**, Mean (SD)		
	method	**K**	**MCE,%**	**Minutes**	**K**	**MCE,%**	**Minutes**	**K**	**MCE,%**	**Minutes**
σ^2=0.01, τ^2=0.01	EPEM	4.2 (0.5)	2.4 (4.6)	0.6 (0.2)	4.1 (0.3)	0.8 (2.7)	2.4 (0.7)	4.2 (0.4)	2.3 (4.3)	249 (53.3)
	I-EPEM	4.3 (0.5)	2.5 (4.5)	0.6 (0.3)	4.1 (0.2)	0.5 (2.3)	1.3 (0.4)	4.2 (0.4)	1.1 (3.4)	17.0 (5.5)
σ^2=0.01, τ^2=0.1	EPEM	4.2 (0.5)	1.8 (3.8)	0.8 (0.4)	4.1 (0.4)	1.0 (2.9)	3.2 (1.0)	4.0 (0.2)	0.4 (1.9)	236 (68.6)
	I-EPEM	4.2 (0.4)	1.7 (3.9)	0.5 (0.2)	4.1 (0.2)	0.6 (2.5)	1.3 (0.5)	4.1 (0.3)	0.5 (2.4)	15.2 (6.2)
σ^2=0.01, τ^2=0.4	EPEM	3.7 (0.5)	7.4 (11.5)	1.1 (0.5)	3.7 (0.5)	7.1 (11.3)	3.8 (1.1)	3.8 (0.4)	4.4 (9.7)	247 (80.8)
	I-EPEM	3.5 (0.5)	11.4 (12.4)	0.5 (0.2)	3.6 (0.5)	9.7 (12.1)	1.2 (0.6)	3.9 (0.3)	2.5 (7.4)	14.0 (5.8)
σ^2=0.1, τ^2=0.01	EPEM	4.2 (0.4)	1.6 (3.9)	0.6 (0.2)	4.0 (0.1)	0.1 (1.1)	2.7 (0.8)	4.1 (0.3)	0.9 (2.8)	243 (79.3)
	I-EPEM	4.1 (0.4)	1.4 (3.6)	0.6 (0.2)	4.0 (0.1)	0.2 (1.3)	1.3 (0.6)	4.1 (0.3)	0.7 (2.4)	18.0 (7.4)
σ^2=0.1, τ^2=0.1	EPEM	4.3 (0.5)	2.5 (4.5)	0.6 (0.3)	4.1 (0.3)	1.0 (2.9)	2.9 (1.1)	4.2 (0.4)	1.8 (3.8)	239 (50.7)
	I-EPEM	4.2 (0.4)	2.3 (4.3)	0.6 (0.2)	4.1 (0.2)	0.5 (2.1)	1.3 (0.6)	4.1 (0.3)	1.1 (3.1)	16.1 (6.4)
σ^2=0.1, τ^2=0.4	EPEM	4.2 (0.6)	4.9 (7.8)	1.0 (0.6)	4.3 (0.5)	3.1 (5.5)	3.9 (1.6)	4.1 (0.3)	1.1 (3.0)	267 (75.5)
	I-EPEM	4.2 (0.6)	5.7 (8.6)	0.6 (0.2)	4.1 (0.4)	2.3 (5.3)	1.2 (0.6)	4.2 (0.4)	1.6 (3.7)	15.6 (6.1)
σ^2=0.4, τ^2=0.01	EPEM	4.1 (0.5)	4.4 (7.8)	0.8 (0.4)	4.0 (0.2)	0.6 (1.6)	3.4 (1.0)	4.1 (0.3)	1.1 (2.9)	260 (82.1)
	I-EPEM	4.0 (0.5)	4.6 (8.4)	0.6 (0.2)	4.0 (0.1)	0.7 (1.4)	1.4 (0.6)	4.1 (0.3)	1.2 (2.7)	17.8 (7.5)
σ^2=0.4, τ^2=0.1	EPEM	3.5 (0.6)	17.0 (11.0)	0.8 (0.3)	3.6 (0.5)	11.7 (11.7)	3.4 (1.0)	4.1 (0.3)	1.9 (2.2)	256 (81.0)
	I-EPEM	3.4 (0.5)	19.6 (9.7)	0.5 (0.2)	3.4 (0.5)	15.4 (11.3)	1.1 (0.4)	4.1 (0.3)	2.4 (2.8)	17.2 (6.9)
σ^2=0.4, τ^2=0.4	EPEM	3.3 (0.5)	24.5 (3.3)	0.7 (0.3)	3.1 (0.3)	24.9 (2.0)	3.0 (1.0)	3.5 (0.6)	21.2 (8.2)	258 (80.4)
	I-EPEM	3.2 (0.4)	24.7 (2.7)	0.5 (0.2)	3.1 (0.2)	25.0 (1.3)	1.0 (0.3)	4.2 (0.5)	5.9 (4.6)	16.1 (5.5)

B. No replicate-specific RE (Eq2r0)

		50 genes/cluster, Mean (SD)			**125 genes/cluster**, Mean (SD)			**1250 genes/cluster**, Mean (SD)		
	method	**K**	**MCE,%**	**Minutes**	**K**	**MCE,%**	**Minutes**	**K**	**MCE,%**	**Minutes**
σ^2=0.01, τ^2=0.01	EPEM	4.3 (0.5)	2.9 (5.5)	0.4 (0.1)	4.0 (0.2)	0.3 (1.7)	1.7 (0.5)	4.1 (0.3)	0.7 (2.6)	200 (67.2)
	I-EPEM	4.2 (0.5)	2.2 (4.5)	0.3 (0.1)	4.0 (0.1)	0.1 (1.1)	0.8 (0.2)	4.2 (0.4)	1.2 (3.1)	14.4 (4.8)
σ^2=0.01, τ^2=0.1	EPEM	4.5 (0.7)	4.7 (6.4)	0.9 (0.4)	4.1 (0.4)	1.2 (3.6)	3.2 (0.9)	4.4 (0.7)	3.5 (5.8)	223 (86.1)
	I-EPEM	4.5 (0.6)	4.4 (6.3)	0.3 (0.1)	4.1 (0.4)	1.1 (3.7)	0.9 (0.3)	5.6 (0.9)	**13.3 (8.5)**	15.7 (5.4)
σ^2=0.01, τ^2=0.4	EPEM	4.4 (0.6)	3.5 (5.6)	2.9 (1.2)	4.2 (0.4)	1.1 (3.4)	9.2 (2.9)	4.6 (0.8)	5.5 (7.1)	266 (99.0)
	I-EPEM	4.4 (0.6)	4.1 (6.1)	0.3 (0.1)	4.1 (0.4)	1.4 (3.9)	0.8 (0.2)	5.9 (0.9)	**17.6 (8.6)**	14.0 (5.5)
σ^2=0.1, τ^2=0.01	EPEM	4.2 (0.4)	1.8 (4.4)	0.4 (0.2)	4.0 (0.2)	0.2 (1.4)	1.7 (0.5)	4.1 (0.3)	1.0 (4.2)	158 (57.9)
	I-EPEM	4.2 (0.4)	1.7 (4.0)	0.3 (0.1)	4.0 (0.1)	0.1 (1.1)	0.7 (0.2)	4.1 (0.3)	0.6 (2.2)	12.2 (4.1)
σ^2=0.1, τ^2=0.1	EPEM	4.3 (0.6)	3.4 (6.3)	0.5 (0.2)	4.0 (0.2)	0.7 (3.6)	2.0 (0.6)	4.1 (0.3)	0.9 (2.9)	184 (71.1)
	I-EPEM	4.2 (0.5)	2.8 (5.7)	0.3 (0.1)	4.0 (0.2)	0.3 (1.9)	0.8 (0.2)	4.1 (0.4)	1.0 (2.8)	12.6 (4.7)
σ^2=0.1, τ^2=0.4	EPEM	4.2 (0.7)	7.0 (9.4)	0.8 (0.3)	4.0 (0.4)	2.9 (7.4)	2.8 (0.8)	4.2 (0.6)	3.3 (6.6)	210 (56.6)
	I-EPEM	4.2 (0.7)	6.4 (9.1)	0.4 (0.2)	4.0 (0.3)	0.8 (3.4)	0.9 (0.3)	4.7 (0.9)	6.2 (7.4)	15.0 (6.3)
σ^2=0.4, τ^2=0.01	EPEM	4.0 (0.6)	6.4 (9.7)	0.5 (0.2)	4.0 (0.2)	0.8 (2.6)	2.0 (0.6)	4.1 (0.3)	1.1 (3.3)	177 (43.9)
	I-EPEM	4.0 (0.6)	6.7 (9.8)	0.3 (0.1)	4.0 (0.1)	0.7 (1.4)	0.7 (0.2)	4.0 (0.2)	0.5 (1.4)	11.1 (3.9)
σ^2=0.4, τ^2=0.1	EPEM	3.5 (0.6)	18.5 (10.8)	0.5 (0.2)	3.3 (0.5)	17.4 (11.1)	2.1 (0.6)	4.1 (0.3)	2.5 (4.3)	185 (45.0)
	I-EPEM	3.4 (0.6)	19.7 (9.7)	0.3 (0.1)	3.4 (0.5)	17.3 (10.9)	0.8 (0.2)	4.0 (0.2)	1.7 (1.7)	11.1 (5.0)
σ^2=0.4, τ^2=0.4	EPEM	3.2 (0.6)	25.4 (7.8)	0.6 (0.3)	3.0 (0.3)	24.8 (4.7)	2.1 (0.7)	3.9 (0.6)	9.7 (9.7)	196 (64.3)
	I-EPEM	3.2 (0.5)	24.9 (5.9)	0.3 (0.1)	3.0 (0.2)	24.9 (1.9)	0.7 (0.2)	4.0 (0.5)	7.7 (6.9)	13.0 (5.2)

results were similar regardless of method used (EPEM or I-EPEM) except when cluster-size was large (1250 genes per cluster) and between- and within-replicate specific RE was very high (0.4). EPEM performed much poorly with a MCE of 21.2% compared to 5.9% for I-EPEM. In this scenario, using I-EPEM to pre-group the genes helped the clustering algorithm in obtaining more accurate results.

However, when we ignored the replicate-specific RE (Eq2r0; (3.2) with p=2, q=2) and sample size is large and $\sigma_k^2 < \tau_k^2$, I-EPEM performed with a decrease in accuracy compared to EPEM (for example, $\sigma_k^2 = 0.01, \tau_k^2 = 0.4$ resulted in a MCE of 17.6% for I-EPEM compared to 5.5% for EPEM). In this case, the between-replicate error is much higher than the within-replicate error, and not accounting for the additional variability between-replicates in Eq2r0 results in a decrease in accuracy for I-EPEM. In fact, an elevation of MCE for EPEM is also observed (from 0.5% to 3.5% for $\sigma_k^2 = 0.01, \tau_k^2 = 0.1$). Eq2r0 is not able to handle the additional between-replicate variability and therefore results in more predicted clusters (>4.4). The results are further attenuated using I-EPEM when sample size is large (1250 genes per cluster). The results are not as obvious for smaller cluster sizes (50 or 125) because the initialization for I-EPEM and EPEM is much closer (I-EPEM initializes with 125 groups versus 200 or 500 for 50 genes/cluster or 125 genes/cluster for EPEM).

3.3.4 Application to fracture-healing study

3.3.4.1 Polynomial order selection

From SVD, the top 4 eigenvectors were obtained, which explained 91% of the variability of the data. Model selection using BIC was used by varying p from 0 to 4 to determine the optimal p for each of the L eigenvectors. From SVD, we chose a

fourth order fixed effect polynomial (p=4). Figure 3.6 shows the top four eigenvectors of the gene-expression data matrix for each strain, which explain more than 90% of the variance. A random effect (RE) order of q=2 was chosen to minimize over-parameterization of our model.

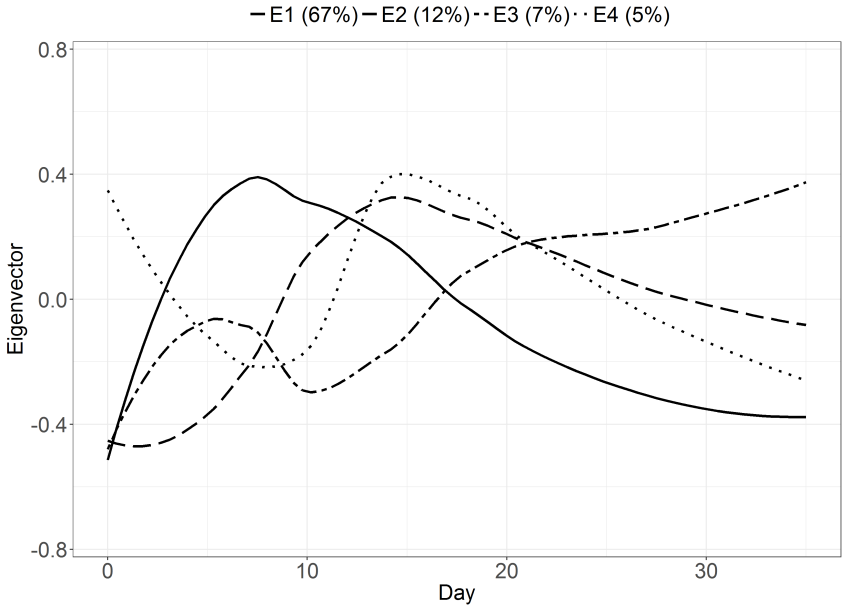

Figure 3.6: Line plots of the top four eigenvectors obtained from singular value decomposition of the fracture healing gene-expression data matrix.

3.3.4.2 Cluster results

I-EPEM

Using the modified initialization approach, 20 clusters are obtained from the data (Figure 3.7). Parameter estimates for each cluster are given in Supplementary Table B.6. Total run-time was 20 hours compared to the 71 hours for the original EPEM algorithm. Clusters 1 through 10 correspond to clusters that have an initial increasing trend. Clusters 13 through 20 correspond to clusters with an initial de-

Table 3.5: Confusion matrix of cluster results from the entropy penalized EM clustering algorithm (EPEM) from Chapter 2 versus the modified initialization EPEM (I-EPEM).

		I-EPEM																			
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**
EPEM	**1**	545	0	1	27	100	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	2	149	865	1	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	3	0	27	477	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	4	6	0	5	690	6	9	14	0	0	0	0	0	0	0	0	0	0	0	0	0
	5	0	0	0	0	349	0	5	0	59	0	0	0	0	0	0	0	0	0	0	0
	6	0	0	0	14	0	772	11	21	0	10	0	0	0	0	0	0	0	0	0	0
	7	0	0	0	0	0	0	639	31	13	15	75	0	0	1	0	0	0	0	0	0
	8	0	0	0	92	97	1	23	205	6	0	0	0	0	0	0	0	0	0	0	0
	9	0	0	0	0	0	0	5	3	310	0	53	1	0	102	0	0	0	0	0	0
	10	0	0	0	0	0	2	0	0	0	595	5	0	0	0	2	0	0	0	0	0
	11	0	0	0	0	0	2	58	321	2	133	635	12	0	0	0	0	0	0	0	0
	12	0	0	0	0	0	0	0	7	0	0	3	165	0	0	0	0	0	0	0	0
	13	0	0	0	0	0	0	0	0	0	1	2702	45	1824	65	13	0	0	0	0	0
	14	0	0	0	0	0	0	0	0	0	0	0	0	1822	0	9	485	1	0	0	0
	15	0	0	0	0	0	0	0	0	0	4	4	6	29	1	635	38	2	0	0	0
	16	0	0	0	0	0	0	0	0	0	0	0	0	50	264	1	402	0	0	0	0
	17	0	0	0	0	0	0	0	0	0	0	0	82	69	0	17	293	56	0	0	3
	18	0	0	0	0	0	0	0	0	0	0	0	0	0	0	13	15	534	8	0	0
	19	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	365	0	0
	20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	630	244
	21	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	7	59	856	63
	22	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	476	106	1	0	1100

creasing trend. Clusters 11 and 12 correspond to genes with flat trends, which can also be evidenced by the small estimated β parameter estimates. In comparing the cluster labels between EPEM and I-EPEM, the ARI is 0.52, suggesting moderate agreement between the two algorithms. A confusion matrix of the cluster labels is given in Table 3.5. A majority of the off-diagonals are zero, suggesting similarity of the two partitions. However, there were many more predicted clusters using EPEM resulting in some clusters from I-EPEM being split between two clusters from EPEM. For example, cluster 11 from I-EPEM was split between 2 clusters for EPEM (11 and 13). Some inconsistencies in how the clusters are being split or defined is occurring between I-EPEM and EPEM, particularly for curves with flatter trajectories.

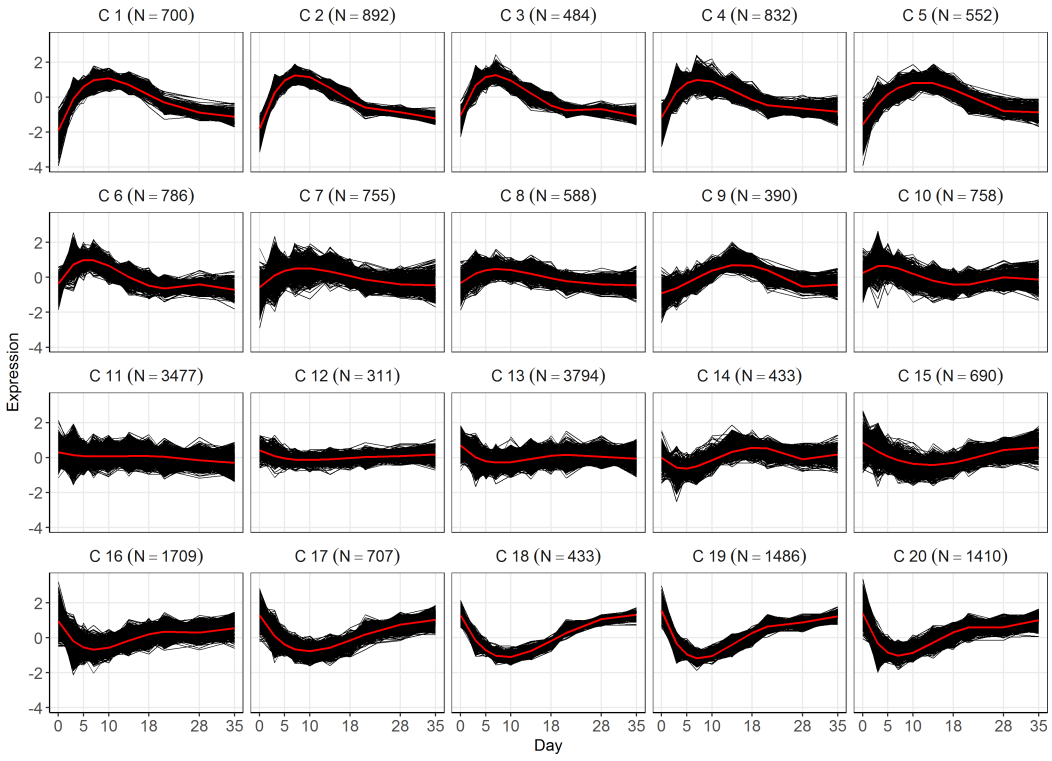

Figure 3.7: Clustering results from fracture healing microarray study with the I-EPEM (Eq2r1; (3.1) with p=4, q=2, strain- (replicate) specific RE). Each plot represents temporal gene-expression profiles clustered into the same group. Total run time was 20 hours. (C: Cluster)

S-EPEM

Using the split clustering approach, 19 clusters are obtained from the data (Figure 3.8). Total run-time was 15 hours, which is lower than I-EPEM, but not greatly so. However, a drastic reduction is similarly observed when compared to the original EPEM algorithm, which took 71 hours to run. The number of genes put into groups with polynomial orders 1, 2, 3 or 4 were 7,259, 1,518, 4,316 and 8,094, respectively. The groups for p=1 and 4 took 7 and 6 hours to run, respectively, due to its large size. Clusters 1 through 8 correspond to clusters that have an initial increasing trend. Clusters 12 through 19 correspond to clusters with an initial decreasing trend. Clusters 9 through 11 correspond to genes with flat or linear trends.

The estimated parameters are given in Supplementary Table B.7. In comparing the cluster labels between EPEM and S-EPEM, the ARI is 0.44, suggesting moderate agreement between the two algorithms, however, the agreement with I-EPEM is slightly higher (0.52). Similarly, a confusion matrix of the cluster labels is given in Table 3.6. The number of off-diagonal entries is slightly higher than for I-EPEM, which can be explained by the additional misclassification of curves into the four groups. Once they have been separated into the four groups, the clusters assigned are restricted to the ones from the independent clustering runs.

3.4 DISCUSSION

In this chapter, we assessed different model selection methods (AIC, BIC, LOOCV, and 10-fold CV) to obtain the optimal fixed effect order to sufficiently represent the data, which would be used in the clustering algorithm. We also assessed two modified versions of the original EPEM algorithm in an attempt to decrease the computational burden in high-dimensional data.

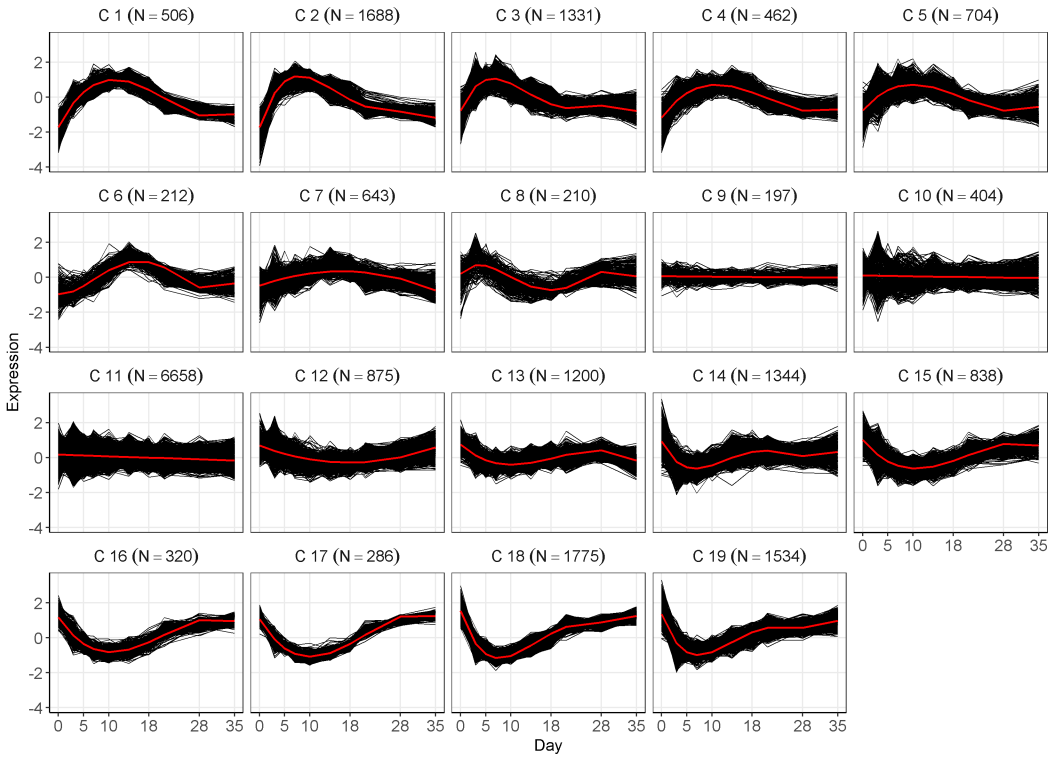

Figure 3.8: Clustering results from fracture healing microarray study with the split entropy penalized EM-algorithm Eq2r1 (p=4, q=2, strain- (replicate) specific RE). Each plot represents temporal gene-expression profiles clustered into the same group. Total run time was 15 hours. (C: Cluster)

Table 3.6: Confusion matrix of cluster results from the entropy penalized EM clustering algorithm (EPEM) from Chapter 2 versus the split EPEM (S-EPEM).

		S-EPEM																		
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**
EPEM	**1**	368	291	3	11	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	2	0	1020	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	3	0	293	215	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	4	11	84	420	31	184	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	5	126	0	0	166	63	53	5	0	0	0	0	0	0	0	0	0	0	0	0
	6	0	0	642	10	54	0	1	8	0	17	96	0	0	0	0	0	0	0	0
	7	0	0	22	0	292	3	149	11	0	79	218	0	0	0	0	0	0	0	0
	8	1	0	18	238	72	0	23	0	8	0	64	0	0	0	0	0	0	0	0
	9	0	0	0	6	7	152	241	0	1	4	60	0	0	3	0	0	0	0	0
	10	0	0	10	0	5	0	0	141	0	102	229	117	0	0	0	0	0	0	0
	11	0	0	1	0	27	1	81	1	18	0	1011	23	0	0	0	0	0	0	0
	12	0	0	0	0	0	0	0	0	151	0	13	11	0	0	0	0	0	0	0
	13	0	0	0	0	0	2	142	1	3	51	4014	173	200	64	0	0	0	0	0
	14	0	0	0	0	0	0	0	0	0	23	533	91	914	649	107	0	0	0	0
	15	0	0	0	0	0	0	0	41	0	34	42	390	36	4	172	0	0	0	0
	16	0	0	0	0	0	1	1	0	0	93	206	13	5	371	27	0	0	0	0
	17	0	0	0	0	0	0	0	0	16	1	170	56	45	45	127	54	0	0	6
	18	0	0	0	0	0	0	0	7	0	0	0	1	0	10	265	202	47	2	36
	19	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	232	127	5
	20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	695	179
	21	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	21	7	931	26
	22	0	0	0	0	0	0	0	0	0	0	2	0	0	198	140	41	0	20	1282

Of the four model selection method's we considered, BIC performed with the highest accuracy in predicting the true order of each of the simulated gene expression profiles. This is to be expected as the BIC was developed for situations where the assumption is that the true existence of a model that is in the scope of all models considered in the selection, where the selected model converges to the true data generating model (Schwarz et al., 1978) (Heinze et al., 2018). This could explain why BIC performed with much higher precision when compared to AIC or CV.

However, on a gene-by-gene level, for all datasets, the optimal order was always found to be at the maximum of the range of orders we considered in model selection (i.e. 4) as the dataset consisted of thousands of genes. Alternatively, using SVD to extract the top subspaces (or eigenvectors), which model selection was performed on, we found that, the optimal order, p, could be estimated and

used for EPEM. Furthermore, investigating the effect of our fixed effect order, p, on EPEM suggested that over-specification of the mixed effects model (with p=4) is preferable to under-specification (with p=2).

Overall, the EPEM algorithm worked well, however, in high-dimensional data with thousands of genes or many potential clusters, the computational burden is high. The first few iterations of the algorithm are costly as it initializes the number of clusters with the number of genes. For example, if we had 20,000 gene expression profiles, the initialization step must determine parameters for each of the 20,000 clusters.

Overall, the EPEM algorithm worked well, however, in high-dimensional data with thousands of genes or many potential clusters, the computational burden is high. The first few iterations of the algorithm are costly as it initializes the number of clusters with the number of genes. For example, if we had 20,000 gene expression profiles, the initialization step must determine parameters for each of the 20,000 clusters.

One approach to solving this issue was to split the data into subgroups that we can cluster separately. However, the split must be done in such a way to minimize initial misclassification's into each of the subgroups. As a result, we defined a split EPEM algorithm to cluster the data in groups based on the predicted polynomial order as determined by model selection using BIC. While the approach successfully decreased the convergence time of the algorithm, the MCE was slightly higher compared to our original approach. However, for high-dimensional data, this drastic reduction in computation time could be worth the slight increase in MCE.

Alternatively, in an effort to solve the initial burden in the first few iterations of

the EPEM, a modified initialization approach was proposed (I-EPEM). Pre-grouping the genes based on their predicted fixed-effects polynomial regression coefficients allowed a way to decrease the initial cluster number used in the algorithm. I-EPEM decreased the computation time compared to EPEM by less than half (20 hours versus 71 hours). In data with a large number of clusters (K=24), I-EPEM performed well, if not better than the original EPEM algorithm.

Additionally, in mixed effects models accounting for the replicate-specific variability, the accuracy in predicted class labels was similar to EPEM in the majority of scenarios considered. However, when sample size was high and the replicate-specific variability was high (τ_k^2=0.4) with a much lower within-replicate variability ($\sigma = 0.01$) and the mixed effects model did not account for this variability (Eq2r0: No Rep), we found that I-EPEM performed with much higher error, suggesting the importance in accounting for this variability in the initialization step due to the large reduction in initial clusters. As the S-EPEM algorithm works with the same methods as the EPEM, but with a reduced sample size, we can assume that the behavior is similar to EPEM in the scenarios considered in Table 3.4, but with a slight increase attributed to any misclassification of genes into an incorrect split. However, some investigators my sacrifice some precision in cluster results for a large reduction in run time (15 hours versus 71 hours between S-EPEM and EPEM, respectively). However, if the clustering algorithm is run in parallel for each of the subsets obtained from S-EPEM, the runtime would be even further reduced to 8 hours.

Generally, for extremely large datasets (>20,000 genes) typical in gene expression datasets, waiting 3 days or more for results may not be feasible. In this case, using I-EPEM can drastically decrease computation time. However, in smaller

datasets (especially when N$< 5^p+1$), EPEM should be used to cluster the data. Additionally, the difference in computation time between S-EPEM and I-EPEM was small because each run of the clustering algorithm for each of the three splits were not run in parallel. In the fracture healing dataset, if clustering for each split was done in parallel, then the run-time would be <7 hours saving even more computation time. However, results would be expected to have a higher misclassification error.

In data with more time-points, future work could involve extending the I-EPEM model to incorporate spline models (as opposed to polynomial models). Splines have been shown to perform with slightly higher accuracy in previous works Chamroukhi (2015) (Gaffney & Smyth, 1999).

While it is important for a clustering algorithm to produce accurate results, if these results are not easily obtainable in terms of computational time, its popularity will be diminished. Using a naÃŕve hierarchical grouping of our genes based on easily obtainable estimates (fixed effects coefficients), the usability of our original algorithm is much improved with a much lower computational burden without a sacrifice in accuracy in gene expression data.

CHAPTER 4

Evaluation of differences in temporal gene-expression patterns by mouse strain

4.1 BACKGROUND

A bone fracture is a medical condition where the continuity of the bone is broken, commonly resulting from high force impact or stress. However, medical conditions which weaken the bones (i.e. osteoporosis) can also cause fractures (Praemer et al., 1992) (NIH, 2000) (US, 2004). Fracture healing is a complex trait, and therefore it is necessary to determine how polygenic networks affect cellular activity during the healing process. Identification of variations in genetic background related to differences in the rate of fracture healing is important in defining when an individual can resume weight bearing activities in their day-to-day life. Information on strain differences (each with different genetic backgrounds) is an important part of understanding and defining individuals with increased risk for complications that can arise during the fracture healing process.

In a previous study on three inbred mouse strains, A/J (AJ), C57Bl/6J (B6) and C3H/HeJ (C3H), different strains of mice were found to have different rates of fracture healing based on regains in strength and stiffness of the bone. Both AJ and B6 strains showed faster healing than the C3H strain, particularly in relation to the length of periods in chondrocyte maturation (Jepsen et al., 2008). B6 exhibited the longest time in each, whereas C3H exhibited the shortest. The study was able to show that variations in skeletal stem cell lineage differentiation existed between strains, and these differences affected the rate of bone fracture healing. To be able to differentiate sets of genes that exhibit this variability across strain would be an important part of understanding the nature of these differences.

To study complex biologic regulatory systems, time-course gene expression data can be obtained from microarrays. Microarrays have been used to simultaneously measure gene expression levels of thousands of genes. The large number of genes and the complexity of the biology, makes clustering analysis a useful and popular tool to analyze such data. Clustering genes with similar temporal expression profiles can be used to identify sets of genes that may be regulated by the same biological mechanism. Gaussian mixture models are a popular approach to cluster analysis, where each gene is assumed to have originated from one component of the mixture model (Fraley et al., 2012). Different extensions to the mixture model have been considered including mixtures of polynomial fixed- and mixed-effects models (Gaffney & Smyth, 1999) (Celeux et al., 2005) as well as extensions to penalized likelihood estimation of the model parameters (Yang et al., 2012) (Chamroukhi, 2015) (Lu et al., 2018).

In previous chapters, incorporation of gene- and replicate-specific random-effects (REs) to an entropy-penalized mixture of polynomial regression model using the EM-Algorithm (EPEM) (Lu et al., 2018) resulted in cluster labels with a lower misclassification error (MCE) than a fixed-effects (FE) only model (Chamroukhi, 2015). However, the complexity of the process grows quickly with the number of genes in the dataset due to the initialization of the algorithm. A modification to the initialization step of EPEM (I-EPEM) was comparable to the original approach with a much decreased computation time and similar degree of accuracy in predicted cluster labels.

In previous applications of the bone-healing data, the clustering was done accounting for strain-specific variability. However, the actual effect of strain on the temporal gene-expression patterns was not determined, it was merely accounted

for. In order to separate out the strain-specific effects on gene-expression profiles, a common approach is to independently cluster the data for each strain and compare results between each set of cluster labels. However, as the clustering was conducted independently, cluster 1 for strain AJ may not be the same as cluster 1 from strain B6, and an additional step in the analysis is to re-match cluster labels between the strains. Issues may arise because the same number of clusters may not be obtained for each strain.

As a result, a new analysis strategy was proposed in this chapter using I-EPEM to compare patterns of temporal gene-expression between different mice-strains. In this new strategy, instead of independent cluster runs for each strain, one run of the clustering algorithm is conducted where expression-profiles for each strain were treated as separate objects to cluster. Sets of genes clustered into different groups across strain can be easily determined as well as sets of genes with the same temporal pattern across strain can also be determined, which is an important part of understanding the underlying biological process of healing. Gene sets with vastly different patterns (i.e. no temporal trend versus an increasing or decreasing trend or one increasing versus one decreasing trend) across strain can also be identified. Performance of this strategy was assessed with a simulation study.

For genes with similar overall patterns (i.e. both increasing, but with different magnitude or timing of maximum expression over the bone healing interval), we borrowed methods from traditional pharmacokinetics (PK) non-compartmental methods (Chen et al., 2001) (Food et al., 2014) to compare differences in the properties of the temporal curve across strain. Traditionally, PK studies seek to quantify the time course of drug absorption, distribution, metabolism and excretion of the drug in the body. To test if two drugs have the same efficacy or toxicity, a bioequiv-

alence study can be conducted to determine if two blood concentration-time profiles are "equivalent". They compare parameters derived from the concentration-time curve such as the area under the curve (AUC: measure of absorption rate), maximum concentration (Cmax: a certain level of concentration to guarantee therapeutic effect), and time of maximum concentration (Tmax; how quickly a drug reaches peak concentration). We extended the use of these measures to temporal gene-expression curves to conduct pairwise differences between the three strains to find sets of genes with differences in the overall magnitude or timing of maximum or minimum gene expression. These vertical or horizontal shifts in gene-expression may help to explain the different rates of fracture healing observed between mice strains.

Finally, an enrichment-analysis using KEGG pathways (Kyoto Encyclopedia of Genes and Genomes (KEGG) Pathways; http://www.genome.jp/kegg/ pathway.html) was conducted on selected sets of genes that exhibited significant horizontal or vertical shifts. Identification of these time-shifted genes could help to explain the different rates of bone fracture healing previously seen in the three strains of mice. For example, Jepsen et. al found that the slower healing mice C3H, had an earlier induction of osteogenesis compared to AJ mice. Other studies also found different durations of osteogenesis or chondrogenesis between the three strains (Grimes et al., 2011). Therefore, a gene-enrichment analysis can give us vital information for genes that exhibited an earlier time to maximum and whether or not they were over-represented in any osteogenesis related pathways. This cluster analysis is a key component to further our understanding of why strain-specific differences are being observed on rates of bone-fracture healing.

4.2 SIMULATION STUDY OF CORRELATED DATA USING I-EPEM AND EPEM

Treating gene-expression profiles for each strain as separate objects to cluster results in a dataset with three temporal gene-expression profiles per gene. Additional correlation is introduced into the clustering algorithm (correlation between the three temporal gene-expression profiles for each gene).

Therefore, a simulation study was conducted to determine if not accounting for this correlation (and assuming independence of temporal curves from the same gene), negatively affects the accuracy of the predicted cluster labels. Let us assume we have four clusters of genes that are each related to a biological pathway or function. Simulated data was obtained to represent data from the same set of genes across two different strains ($y^{(1)}, y^{(2)}$). Let two of the four clusters of genes have strain-specific variations (represented by different fixed-effects parameters, β_{pk}), whereas the other two do not. One way of introducing correlation into the data is by forcing correlation between the expression measurement for a given gene at a given time-point ($\rho(\epsilon^{(1)}_{trik}, \epsilon^{(2)}_{trik})$. The data are simulated from the following model, where (4.1) and (4.2) correspond to data from strain 1 or strain 2, respectively:

$$y^{(1)}_{trik} = \beta^{(1)}_{0k} + \beta^{(1)}_{1k} x_{tri} + \ldots + \beta^{(1)}_{pk} x^{p}_{tri} + b^{(1)}_{0ik} + b^{(1)}_{1ik} x_{tri} + b^{(1)}_{2ik} x^{2}_{tri} + c^{(1)}_{rik} + \epsilon^{(1)}_{trik} \quad (4.1)$$

$$y^{(2)}_{trik} = \beta^{(2)}_{0k} + \beta^{(2)}_{1k} x_{tri} + \ldots + \beta^{(2)}_{pk} x^{p}_{tri} + b^{(2)}_{0ik} + b^{(2)}_{1ik} x_{tri} + b^{(2)}_{2ik} x^{2}_{tri} + c^{(2)}_{rik} + \epsilon^{(2)}_{trik} \quad (4.2)$$

- β_k=($\beta^{(1)}_k$,$\beta^{(2)}_k$)$'$ $= (\beta^{(1)}_{0k}, \ldots, \beta^{(1)}_{0k}, \beta^{(1)}_{pk}, \ldots, \beta^{(2)}_{pk})$ are the k-th cluster FE coefficients for each strain.

- x_{tri} is t-th time-point for the r-th replicate of the i-th temporal gene expression

profile.

- $b_{ik}^{(1)} = (b_{0ik}^{(1)}, b_{1ik}^{(1)}, b_{2ik}^{(1)})^T \sim N(0, G_k^{(1)})$ and $b_{ik}^{(2)} = (b_{0ik}^{(2)}, b_{1ik}^{(2)}, b_{2ik}^{(2)})^T \sim N(0, G_k^{(2)})$ are the gene-specific REs for the each strain such that $b_{ik} = (b_{ik}^{(1)}, b_{ik}^{(2)})' \sim N\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}G_k^{(1)} & 0\\ 0 & G_k^{(2)}\end{pmatrix}\right)$. G_k is a 3×3 matrix.

- $c_{rik} = (c_{rik}^{(1)}, c_{rik}^{(2)})' \sim N\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}\tau_k^{2(1)} & 0\\ 0 & \tau_k^{2(2)}\end{pmatrix}\right)$ is the replicate-specific RE (between replicate variability) for the s-th strain.

- $\epsilon_{trik} = (\epsilon_{trik}^{(1)}, \epsilon_{trik}^{(2)})' \sim N\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}\sigma_k^{2(1)} & 0.9\sigma_k^{(1)}\sigma_k^{(2)}\\ \rho\sigma_k^{(1)}\sigma_k^{(2)} & \sigma_k^{(2)}\end{pmatrix}\right)$ is the measurement error (within-replicate variability) such that the observations for the same strain at the same time-point are correlated.

- For simplicity, let $G_k^{(1)} = G_k^{(2)}$, $\sigma_k^{2(1)} = \sigma_k^{2(2)}$, $\tau_k^{2(1)} = \tau_k^{2(2)}$, and ρ=0.9.

- Assume that b_{ik}, c_{rik} and ϵ_{trik} are mutually independent and each cluster is allowed to have a different mean vector and covariance matrix.

The parameters used in the simulation are specified in Table 4.1. Var(b_i), correspond to diagonal entries of the G-matrix, whereas covariance between b_i and b_j for ($i \neq j$) is equal to $\rho(b_i, b_j)\sqrt{Var(b_i)}\sqrt{Var(b_j)}$. One iteration of simulated data is shown in Figure 4.1. The data consist of four clusters, where clusters 1 and 2 show obvious shifts in expression as a result of strain 1 or 2 (denoted by a black or grey line, respectively). Cluster 1 shows a shift in maximum expression by one day, and cluster 2 shows a shift in minimum expression by 2 days. Cluster 3 and 4 show no change in expression due to strain. The goal of the simulation study was to determine the impact on accuracy of the predicted cluster labels using I-EPEM or EPEM by introducing correlation into the data.

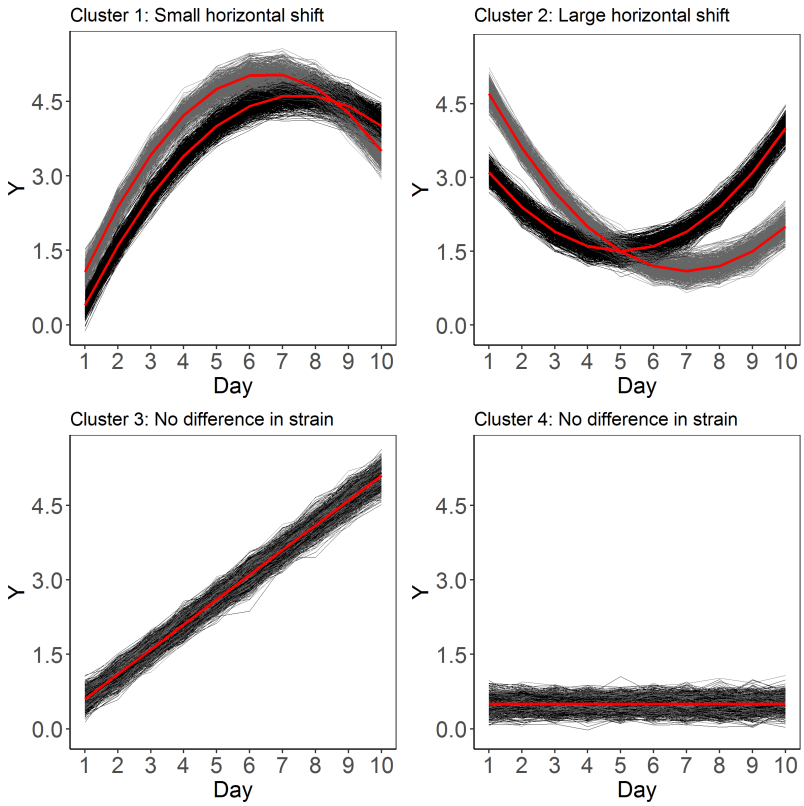

Figure 4.1: Simulation plot with four true clusters (clusters 1 and 2 show a small or large strain-specific horizontal shift in maximum or minimum expression; clusters 3 and 4 show no strain-specific change in expression). Black lines correspond to strain 1 and gray lines correspond to strain 2, and the red line corresponds to the average expression for a given cluster and strain.

Table 4.1: Simulation parameters to obtain datasets used in simulation studies with with equal mixing proportions, 125 genes per cluster, 4 replicates and varying within- (σ_k^2) or between- (τ_k^2) replicate variability.

	Cluster 1		Cluster 2		Cluster 3		Cluster 4	
Parameter	**Strain 1**	**Strain 2**	**Strain 1**	**Strain 2**	**Strain 1**	**Strain 2**	**Strain 1**	**Strain 2**
β_0	-1.0	-0.5	4.0	6.0	0.1	0.1	0.5	0.5
β_1	1.5	1.7	-1	-1.4	0.5	0.5	–	–
β_2	-0.1	-0.13	0.1	0.1	–	–	–	–
Var(b_0)[a]	$0.2\sigma^2$		$0.8\sigma^2$		$0.02\sigma^2$		$0.1\sigma^2$	
Var(b_1)[a]	$0.3\sigma^2$		$0.2\sigma^2$		$0.1\sigma^2$		–	
Var(b_2)[a]	$0.02\sigma^2$		$0.02\sigma^2$		–		–	
σ^2	0.01, 0.10, or 0.40							
τ^2	0.01, 0.10, or 0.40							

[a]Correlation (ρ) between b_i such that $\rho(b_0, b_1) = -0.5$, $\rho(b0, b2) = 0.4$ and $\rho(b1, b2) = -0.9$

The entropy penalized EM algorithm (EPEM; Chapter 2) and the modified initialization EPEM algorithm (I-EPEM) were used to cluster the data using an underlying model with p=2, q=2, with and without a replicate-specific RE ((4.3) and (4.4), respectively). Assuming a true cluster number of 6, the mean (SD) number of clusters (K) and overall misclassification error (MCE) for 1000 iterations is determined.

$$y_{trik} = \beta_{0k} + \beta_{1k}x_{tri} + \beta_{2k}x_{tri}^2 + b_{0ik} + b_{1ik}x_{tri} + b_{2ik}x_{tri}^2 + c_{rik} + \epsilon_{trik} \tag{4.3}$$

$$y_{tik} = \beta_{0k} + \beta_{1k}x_{ti} + \beta_{2k}x_{ti}^2 + b_{0ik} + b_{1ik}x_{ti} + b_{2ik}x_{ti}^2 + \epsilon_{tik} \tag{4.4}$$

4.3 EVALUATION OF STRAIN-SPECIFIC DIFFERENCES IN BONE FRACTURE HEALING

4.3.1 The bone fracture-healing data

The data consist of 3 strains of mice (AJ, B6 and C3H) with gene expression measurements taken over 10 time points (Days 0, 3, 5, 7, 10, 14, 18, 21, 28, and 35). Conducting independent strain-specific clustering accounting for a replicate-specific affect resulted in very low estimated replicate-specific variabilities (<0.01). Furthermore, from our simulation studies in this chapter, scenarios with low between-replicate variability with or without a replicate-specific RE resulted in cluster labels with similar accuracy. Therefore, to simplify the model, the 3 replicates at each time point are averaged so that there are 10 gene expression measurements defining each temporal curve. See Appendix A.1 for more details on how the expression measurements were obtained.

4.3.2 Modified Initialization of Entropy Penalized EM Algorithm

$$y_{tik} = \beta_{0k} + \beta_{1k}x_{ti} + \beta_{3k}x_{ti}^2 + \beta_{3k}x_{ti}^3 + \beta_{4k}x_{ti}^4 + b_{0ik} + b_{1ik}x_{ti} + b_{2ik}x_{ti}^2 + \epsilon_{tik} \qquad (4.5)$$

An underlying model specified by (4.5) is used in the I-EPEM clustering algorithm defined by a fourth order FE polynomial and a second order gene-specific RE polynomial. β_k=(β_{0k},β_{1k},...,β_{4k})′ are the kth cluster FE polynomial coefficients; x_{ti} is t-th time-point for the i-th temporal gene-expression profile (note that i=1, . . ., 21,187×3); b_{ik}=(b_{0ik},b_{1ik},b_{2ik})′ $\sim N(0, G_k)$ is the gene-specific RE where $G_k \in R^{(3\times3)}$; $\epsilon_{tik} \sim N(0, \sigma_k^2)$ is the measurement error (within-replicate variability). Assume b_{ik} and ϵ_{tik} are mutually independent and each cluster is allowed to have a different mean vector and covariance matrix. The 21,187×3 temporal gene expression pro-

files were pre-grouped into 5^4 = 3,125 initial clusters.

4.3.3 Pairwise comparisons for strain

Cross-tabulations of the cluster results for genes from each strain were obtained for each strain pair (AJ versus B6, AJ versus C3H, and B6 versus C3H). Genes in clusters on the diagonal of the cross-tabulation (i.e. cluster 1 for AJ and cluster 1 for B6) represent those that were grouped into the same cluster across strain. Genes off the diagonal represent those that were grouped into a different cluster across strain (i.e. cluster 1 for AJ and cluster 2 for B6). Off diagonal gene groups were separated into six types of trends: (1) both strains with no temporal trend, (2) no temporal trend versus an initial increasing temporal trend, (3) no temporal trend versus an initial decreasing temporal trend, (4) both initial increasing temporal trends grouped into different clusters, (5) both initial decreasing temporal trends grouped into different clusters, (6) one initial increasing and one initial decreasing temporal trend (6) both with no temporal trend. Only groups with at least 100 genes will be considered for practicality in pathway or functional analyses of the gene-sets.

Pairwise comparisons for types (4) and (5) are conducted to determine if any vertical (magnitude of maximum expression) or horizontal (time of maximum expression) shifts occurred, despite similar overall increasing or decreasing patterns. Genes in groups (1), (2), (3) and (6) are not compared in this way as the nature of the difference in the curves does not lend to such a comparison. Given that these groups of genes were grouped into completely different clusters with different patterns (i.e. increasing versus decreasing), we assume that strain differences exist and are detected by the clustering algorithm itself.

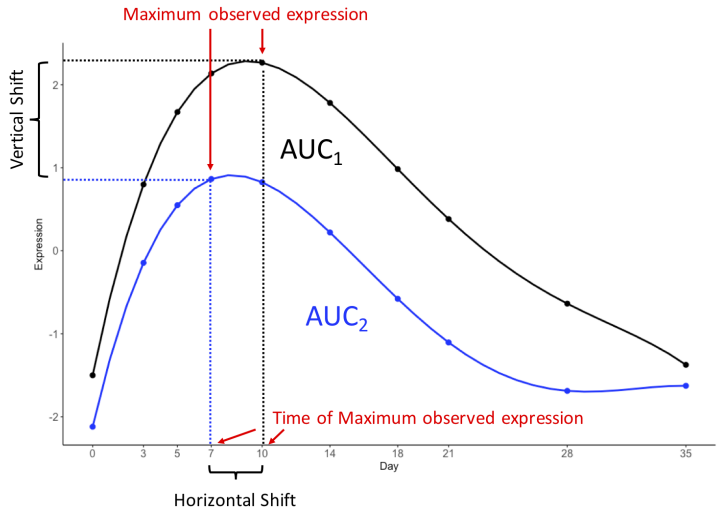

Figure 4.2: Features of the temporal gene-expression curve to be compared between strains.

4.3.3.1 Comparison of features of the temporal gene-expression curves between strain

For genes in groups (4) and (5), both with an overall increasing or decreasing trend, pairwise comparisons of specific features of the temporal expression curves will be conducted. These genes differ by the magnitude or rate of increase/decrease in expression values across strain. The features we consider are maximum and time to maximum expression (for group (4)), minimum and time to minimum expression (for group (5)), and area under the curve (AUC), which is depicted in Figure 4.2.

Maximum and Minimum expression

Minimum (Min) and maximum (Max) observed gene expression values were used to determine genes with different magnitudes of expression. The magnitude of expression can be biologically related to the impact of certain signal transduction pathways (Brivanlou & Darnell, 2002) (Ghandhi et al., 2011). Within each group

of genes, the maximum and minimum expression was determined for each gene and strain (ignoring day 0 and 35). Day 0 and 35 were ignored because we were interested in determining vertical shifts in local minimum or maximum (Figure 4.2) between the two curves. A one-sample t-test of the differences between strain was conducted and mean of the differences (95% confidence intervals) are reported to determine if a significant vertical shift of the temporal curve occurred.

AUC

The area under the curve (AUC) can reflect the total expression seen over the fracture healing process. AUC was determined by using the trapezoidal rule (4.6),where t_i corresponds to the i-th time point and y_i corresponds to the i-th expression measurement for a particular curve. As gene-expression measurements can be positive or negative, the values were shifted by the minimum value so that all measurements are greater than zero. The AUC is calculated for each gene and strain. The mean of the differences (95% confidence interval) between two strains are reported and significance is assessed with a one sample t-test.

$$AUC = 0.5 \sum_{i=1}^{n-1} (t_{i+1} - t_i)(y_{i+1} + y_i) \tag{4.6}$$

Time of maximum and minimum expression

Time to minimum and maximum expression reflect the dynamics of individual gene expression and in many cases where common patterns are observed indicate coordinate control of transcription rates of a group of genes by a common transcription factor (Pedraza & Paulsson, 2007) (Singh & Dennehy, 2014) (Rowicka et al., 2007). The time of maximum or minimum observed expression for a given

gene was defined as the time (ignoring day 0 and 35), of the maximum or minimum observed expression within the healing process (Figure 4.2), respectively. A Wilcoxon signed rank test (WSR) to compare distributions of time at max or min between the two strains are used to determine if a significant horizontal shift has occurred. The WSR tests the null hypothesis that the difference between the pairs follows a symmetric distribution around zero versus the alternative that difference between the pairs does not follow a symmetric distribution around zero. As the WSR test ignores zero differences in the calculation of the test statistic, to obtain a better understanding of the distribution of differences, the percent of negative, positive and no difference are also obtained. Summaries for the median of the differences and the interquartile range ignoring zero differences are also reported.

All p-values are adjusted for multiple correction using Bonferroni correction (455 tests). Significance is assessed at an adjusted α level of 1.1×10^{-4}.

4.3.4 Enriched KEGG pathways for a set of genes

KEGG pathways (Kyoto Encyclopedia of Genes and Genomes (KEGG) Pathways; http://www.genome.jp/kegg/pathway.html) consist of manually drawn pathway maps that represent current knowledge on molecular interaction and reaction networks, for a large selection of organisms (including mice). It consists of information on pathways associated with metabolism, genetic information processing, environmental information processing and cellular processes. A gene-enrichment analysis was conducted to illustrate how the results of the cluster analysis could be used. KEGG pathways are determined where a set of genes are overrepresented, compared to a randomly sampled set of genes using a hypergeometric test with

Table 4.2: Simulation results averaged over 1000 iterations for EPEM or I-EPEM clustering algorithms assuming two different underlying mixed effects models (p=2, q=2 with or without a replicate-specific RE). The number of clusters (K) is reported along with overall and cluster-specific misclassification error (MCE). 1000 genes, 4 replicates, assuming 6 true clusters of genes.

		EPEM		I-EPEM	
Model	Error1	K (SD)	MCE (SD), %	K (SD)	MCE (SD), %
Rep RE	σ^2=0.01, τ^2=0.01	6.1 (0.4)	0.54 (1.91)	6.0 (0.1)	0.04 (0.47)
	σ^2=0.01, τ^2=0.1	6.2 (0.5)	0.92 (2.42)	6.0 (0.2)	0.15 (1.18)
	σ^2=0.01, τ^2=0.4	5.9 (0.3)	0.86 (3.17)	5.9 (0.3)	0.95 (3.31)
	σ^2=0.1, τ^2=0.01	6.0 (0.1)	0.13 (1.31)	6.0 (0.1)	0.05 (0.71)
	σ^2=0.1, τ^2=0.1	6.1 (0.3)	0.43 (1.63)	6.0 (0.1)	0.08 (0.53)
	σ^2=0.1, τ^2=0.4	6.2 (0.6)	1.43 (3.08)	6.0 (0.2)	0.34 (1.48)
	σ^2=0.4, τ^2=0.01	5.1 (0.4)	11.26 (4.49)	5.3 (0.5)	9.55 (6.16)
	σ^2=0.4, τ^2=0.1	5.1 (0.4)	11.86 (4.21)	5.2 (0.5)	10.74 (5.23)
	σ^2=0.4, τ^2=0.4	5.4 (0.6)	9.76 (5.03)	5.2 (0.5)	10.96 (4.53)
No Rep RE	σ^2=0.01, τ^2=0.01	6.0 (0.2)	0.13 (1.22)	6.0 (0.1)	0.16 (1.39)
	σ^2=0.01, τ^2=0.1	6.1 (0.3)	0.66 (2.78)	6.0 (0.3)	0.65 (2.85)
	σ^2=0.01, τ^2=0.4	6.1 (0.2)	0.54 (2.28)	6.1 (0.3)	0.82 (3.12)
	σ^2=0.1, τ^2=0.01	6.0 (0.2)	0.60 (2.75)	6.0 (0.1)	0.19 (1.58)
	σ^2=0.1, τ^2=0.1	5.9 (0.4)	1.72 (4.46)	6.0 (0.2)	0.26 (1.76)
	σ^2=0.1, τ^2=0.4	5.7 (0.5)	4.74 (6.18)	5.9 (0.4)	2.19 (4.86)
	σ^2=0.4, τ^2=0.01	5.3 (0.6)	9.27 (6.59)	5.3 (0.6)	9.49 (6.45)
	σ^2=0.4, τ^2=0.1	5.1 (0.5)	11.20 (5.93)	5.2 (0.5)	10.80 (5.56)
	σ^2=0.4, τ^2=0.4	5.0 (0.5)	13.15 (5.14)	5.0 (0.3)	12.65 (3.62)

Rep: Replicate; RE: Random Effect; EPEM: Entropy penalized EM Algorithm; I-EPEM: Modified Initialization EPEM

the R package KEGGPROFILE() (Zhao et al., 2015). Pathways with p-values <0.05 are reported along with the number of genes associated with it.

4.4 RESULTS

4.4.1 Simulation study of correlated data using I-EPEM and EPEM

Using the original EPEM algorithm (Table 4.2), the clustering algorithm was able to accurately separate the data into the 6 clusters for data with low or high variability (σ_k^2=0.01 or 0.10) using a mixed effects model accounting or ignoring the replicate-

specific variability (overall MCE <1.5% for 6 true clusters). As expected, when the within-replicate variability the error is very high (σ_k^2=0.4), the clustering does not perform as well with a high overall MCE (>9% for 6 true clusters) and around 5 predicted clusters. The clustering algorithm combined the data from the two strains with a small horizontal shift. Similar results are seen using the I-EPEM algorithm.

4.4.2 Overall cluster patterns

From Figure 4.3, the set of genes across all strains were grouped into 18 different clusters. The clusters were ordered by overall pattern. Clusters 1 through 6 showed an initial increase in expression. Clusters 7 and 8 showed a flat pattern with no visual change in expression over time, which can also be evidenced by small estimates of predicted β coefficients (Supplementary Table B.8). Clusters 8 through 18 showed an initial decrease in expression. Additionally, from Figure B.1, the temporal curves from Figure 4.3 are plotted by strain. Each set of 18 clusters for each strain consist of the same 21,187 genes. We see that almost no temporal gene expression profile originating from B6 or C3H was clustered into cluster 2. Genes grouped into clusters 1-6 were considered increasing, genes grouped into clusters 9-18 were considered decreasing, and genes grouped into clusters 7 and 8 were considered flat (no temporal trend).

4.4.3 AJ versus B6 comparison

Table 4.3 is a cross-tabulation of the cluster labels for temporal gene-expression profiles belonging to strain AJ versus B6. The gray on-diagonal entries correspond to genes that were grouped into the same cluster. For instance, 176 genes had

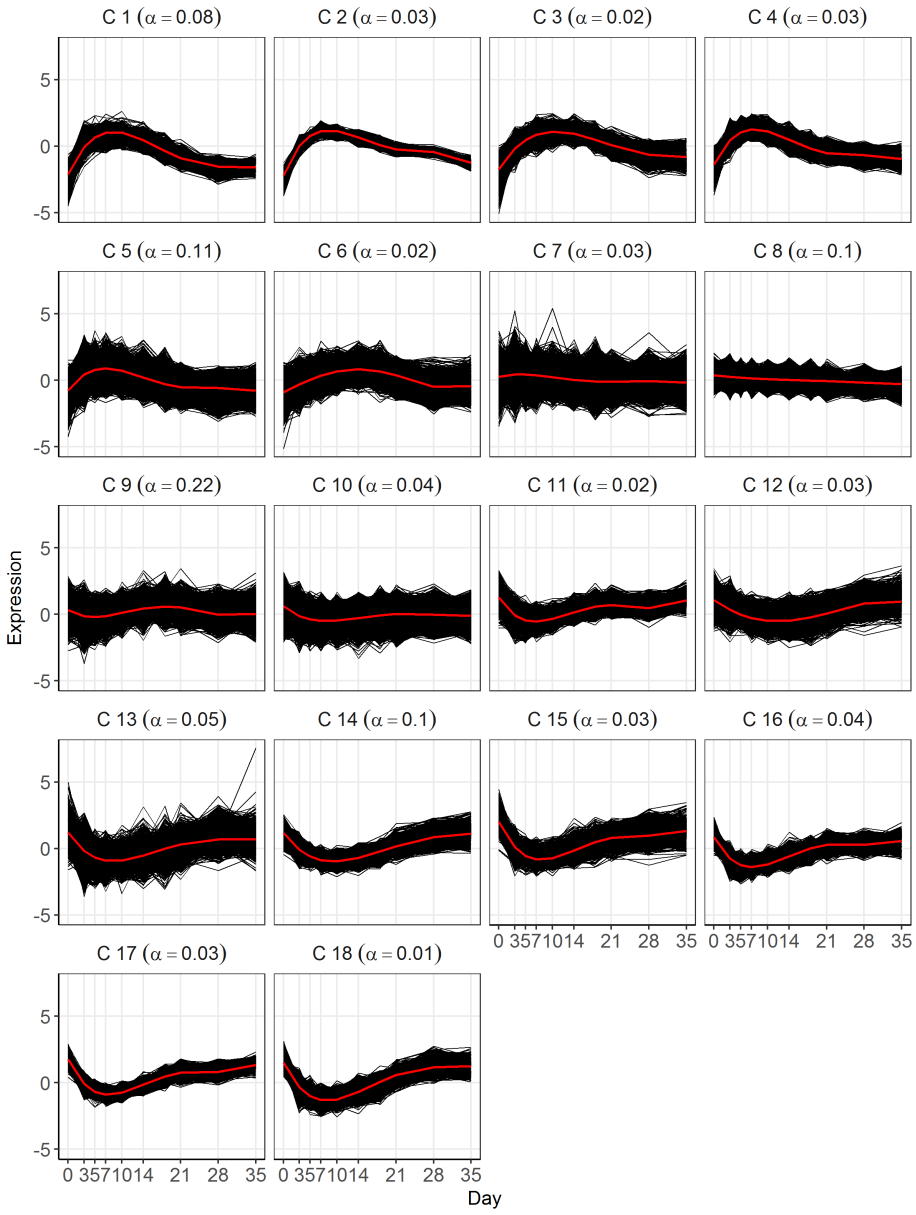

Figure 4.3: Overall cluster results from a modified initialization entropy penalized EM algorithm with 63,561 temporal gene-expression curves across the three strains (AJ, B6 and C3H). α corresponds to the proportion of genes in that particular cluster. (C: cluster)

Table 4.3: Cross-tabulation of AJ versus B6 cluster labels. The shaded cells on-diagonal correspond to genes that were grouped into the same cluster for the AJ and B6 strain.

AJ Strain \ B6 Strain	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
1	3	0	3	12	7	0	0	2	0	0	0	0	0	0	0	0	0	0
2	534	1	9	65	27	1	1	0	0	0	0	0	0	0	0	0	0	0
3	256	0	176	29	270	82	64	5	3	6	0	0	0	0	0	0	0	0
4	385	1	9	353	302	3	23	2	0	2	0	0	0	0	0	0	0	0
5	74	0	41	205	1008	38	266	55	20	36	1	6	4	0	0	0	0	0
6	14	0	55	6	177	283	155	26	93	109	1	2	23	0	0	0	0	0
7	0	0	3	5	308	92	521	167	93	227	4	67	30	5	1	1	0	4
8	5	0	6	4	232	37	482	101	103	204	2	28	54	2	6	2	1	0
9	0	0	4	1	42	117	307	170	316	491	19	31	234	6	12	23	0	8
10	0	0	5	0	157	207	1355	510	1150	1734	118	254	459	48	134	36	9	49
11	0	0	0	0	2	3	17	11	34	164	48	49	423	40	114	88	21	270
12	0	0	0	0	2	0	40	5	6	58	3	87	30	24	4	3	2	9
13	0	0	0	0	0	0	29	19	28	95	20	17	39	12	24	2	2	8
14	0	0	0	0	0	0	8	0	7	43	16	102	27	248	46	7	14	94
15	0	0	0	0	1	0	14	1	10	97	42	21	56	24	107	33	30	95
16	0	0	0	0	1	0	5	3	28	82	27	13	35	23	70	11	13	29
17	0	0	0	0	0	0	1	1	5	21	28	7	54	70	245	35	336	822
18	0	0	0	0	0	0	1	0	0	9	6	2	3	10	15	0	19	25

the same pattern for AJ and B6 and were grouped into the same cluster (cluster 3). Alternatively, off-diagonal entries correspond to genes that were grouped into different clusters. For example, 109 genes that showed an increasing trend for AJ mice (cluster 6) were clustered into a different cluster with a decreasing trend for B6 mice (Cluster 10). In fact, there were 4 groups of genes (Supplementary Figure B.2) that were found such that one strain showed an initial increase of expression compared to the other strain with an initial decrease in expression. Additionally, there were 4 and 7 groups of genes where one strain showed an initial increase or decrease in expression compared to a relatively flat trend in the other strain (Supplementary Figure B.3 and B.4, respectively).

Seven groups of genes were found to be clustered into different clusters both with an overall initial increasing trend. Figure 4.4 plots the mean expression ($\pm$sd) for each strain at each time point for all genes in that group. Table 4.4 summarizes the differences in features between AJ and B6 strain grouped by the type of observed difference. After adjusting for multiple comparisons, no significant strain-specific differences in maximum expression or time to maximum expression

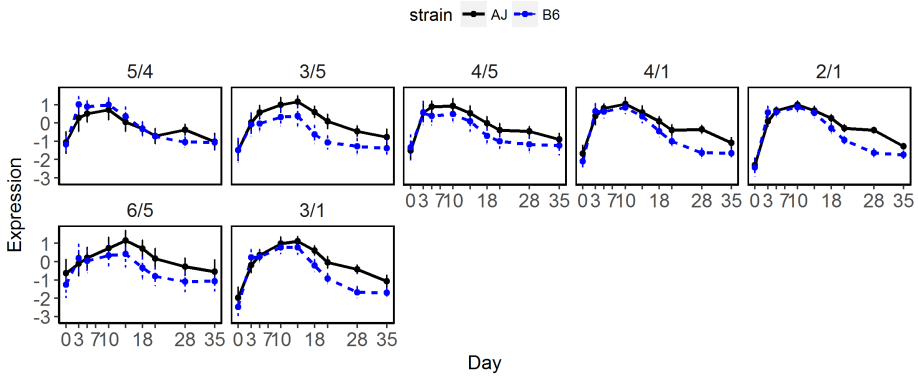

Figure 4.4: Strain-specific mean (SD) time-curve over all genes that were clustered into two different clusters across strain AJ and B6, where both clusters showed an initial increasing trend.

Table 4.4: Summary of features (area under the curve (AUC), maximum expression (Exp), and time at maximum Exp) comparing strain differences (B6-AJ) in temporal gene expression curves clustered into groups with an initial increasing trend.

			Continuous Features		Ordinal Features				
	Cluster			Mean of		Sign of Diff			Median of
Type	AJ/B6	N Genes	Parameter	B6-AJ (95% CI)	Parameter	Negative	Zero	Positive	B6-AJ (IQR)
No horizontal shift	5/4	205	Max Exp	0.14 (0.21, 0.06)	Time at Max Exp	23%	60%	21%	-2 (-4, 3)
No vertical shift			AUC	0.31 (1.91,-1.29)	Time at Max+Min Exp	.	13%	.	
No horizontal shift	3/5	270	Max Exp	-0.69 (-0.61,-0.76)*	Time at Max Exp	30%	49%	16%	-3 (-7, 3)
Vertical shift			AUC	-26.6 (-24.8,-28.5)*	Time at Max+Min Exp	.	26%	.	
	4/5	302	Max Exp	-0.49 (-0.41,-0.56)*	Time at Max Exp	27%	56%	21%	-2 (-4, 3)
			AUC	-17.4 (-15.7,-19.2)*	Time at Max+Min Exp	.	23%	.	
	4/1	385	Max Exp	-0.26 (-0.22,-0.31)*	Time at Max Exp	20%	66%	12%	-3 (-4, 3)
			AUC	-18.6 (-17.5,-19.7)*	Time at Max+Min Exp	.	24%	.	
	2/1	534	Max Exp	-0.17 (-0.15,-0.20)*	Time at Max Exp	11%	76%	17%	3 (-3, 3)
			AUC	-16.0 (-15.3,-16.6)*	Time at Max+Min Exp	.	45%	.	
Horizontal shift	6/5	177	Max Exp	-0.54 (-0.40,-0.68)*	Time at Max Exp	53%	31%	20%	-6 (-8, -3)*
Vertical shift			AUC	-21.8 (-18.2,-25.4)*	Time at Max+Min Exp	.	11%	.	
	3/1	256	Max Exp	-0.29 (-0.24,-0.34)*	Time at Max Exp	21%	66%	13%	-4 (-7, 3)*
			AUC	-20.6 (-19.2,-22.0)*	Time at Max+Min Exp	.	40%	.	

[1] Median (IQR) represents the median ignoring zero-differences.

*p-value < 1.1×10^{-4} (Bonferroni adjustment for 595 total tests conducted). Differences in continuous and ordinal features are tested with a one-sample t-test and wilcoxon signed rank test, respectively.

among genes in AJ/B6 cluster 5/4. Similarly, no significant differences in time to maximum expression was observed in AJ/B6 cluster 3/5, 4/5, 4/1 and 2/1 (more than 49% with zero-difference). However, AJ mice had a significantly higher maximum expression compared to B6 mice, with 3/5 and 2/1 showing the largest and the smallest increase in maximum expression, respectively. Lastly, AJ/B6 clusters 6/5 and 3/1 showed a significant horizontal and vertical time to maximum expression and maximum expression for AJ mice, respectively. AJ mice had a longer time to maximum, but with a higher magnitude of maximum expression compared to B6. The AUC was all significantly different between strains for all groups except for 5/4.

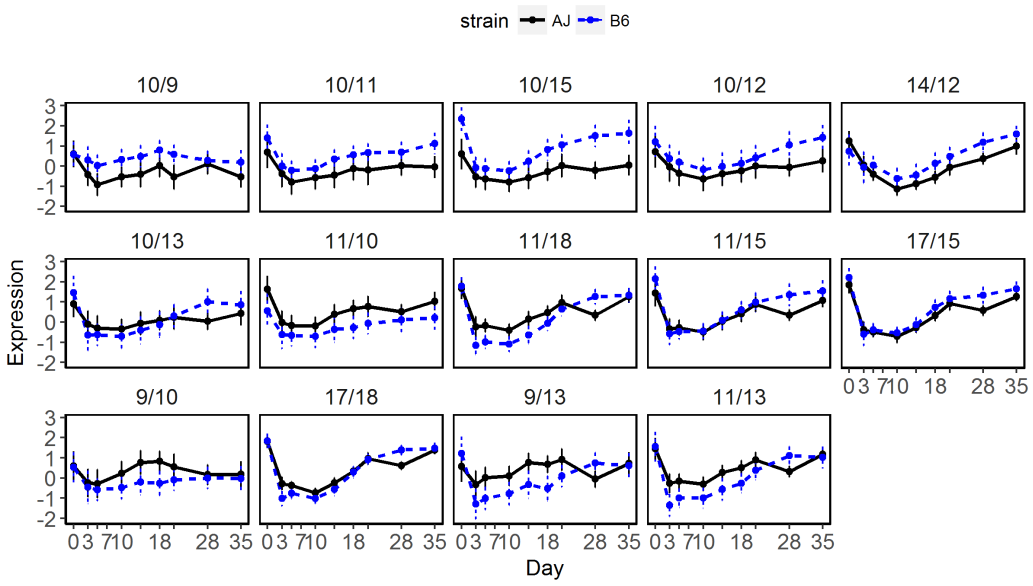

Figure 4.5: Strain-specific mean (SD) time-curve over all genes that were clustered into two different clusters across strain AJ and B6, where both clusters showed an initial decreasing trend.

Table 4.5: Summary of features (area under the curve (AUC), minimum expression (Exp), and time at minimum Exp) comparing strain differences (B6-AJ) in temporal gene expression curves clustered into groups with an initial decreasing trend.

			Continuous Features		Ordinal Features				
	Cluster			Mean of		Sign of Diff			Median of
Type	AJ/B6	N Genes	Parameter	B6-AJ (95% CI)	Parameter	Negative	Zero	Positive	B6-AJ (IQR)
No horizontal shift	10/9	1150	Min Exp	0.87 (0.90, 0.83)*	Time at Min Exp	38%	20%	42%	2 (-5, 7)
Vertical shift			AUC	23.93 (24.90,22.96)*	Time at Max+Min Exp	.	4%	.	
	10/11	118	Min Exp	0.65 (0.76, 0.54)*	Time at Min Exp	50%	18%	32%	-3 (-7, 2)
			AUC	24.10 (27.68,20.51)*	Time at Max+Min Exp	.	7%	.	
	10/15	134	Min Exp	0.54 (0.64, 0.44)*	Time at Min Exp	46%	25%	28%	-3 (-4, 3)
			AUC	37.21 (40.70,33.72)*	Time at Max+Min Exp	.	5%	.	
	10/12	254	Min Exp	0.52 (0.62, 0.42)*	Time at Min Exp	37%	22%	41%	2 (-4, 7)
			AUC	22.18 (25.26,19.11)*	Time at Max+Min Exp	.	4%	.	
	14/12	102	Min Exp	0.47 (0.60, 0.34)*	Time at Min Exp	33%	48%	19%	-3 (-7, 4)
			AUC	17.57 (21.37,13.77)*	Time at Max+Min Exp	.	31%	.	
	10/13	459	Min Exp	-0.39 (-0.32,-0.46)*	Time at Min Exp	42%	30%	27%	-3 (-7, 4)
			AUC	2.79 (4.73, 0.84)	Time at Max+Min Exp	.	4%	.	
	11/10	164	Min Exp	-0.54 (-0.42,-0.66)*	Time at Min Exp	29%	36%	35%	2 (-3, 5)
			AUC	-23.2 (-20.1,-26.3)*	Time at Max+Min Exp	.	10%	.	
	11/18	270	Min Exp	-0.78 (-0.73,-0.84)*	Time at Min Exp	39%	37%	24%	-3 (-4, 4)
			AUC	-8.49 (-7.01,-9.97)*	Time at Max+Min Exp	.	3%	.	
Horizontal shift	11/15	114	Min Exp	-0.11 (-0.03,-0.20)	Time at Min Exp	44%	38%	18%	-3 (-4, 2)*
No vertical shift			AUC	9.88 (12.68, 7.08)*	Time at Max+Min Exp	.	11%	.	
Horizontal shift	17/15	245	Min Exp	0.12 (0.17, 0.07)*	Time at Min Exp	48%	46%	6%	-4 (-4, -3)*
Vertical Shift			AUC	11.16 (12.58, 9.75)*	Time at Max+Min Exp	.	18%	.	
	9/10	491	Min Exp	-0.32 (-0.24,-0.39)*	Time at Min Exp	24%	23%	53%	3 (-2, 9)*
			AUC	-17.9 (-16.1,-19.7)*	Time at Max+Min Exp	.	4%	.	
	17/18	822	Min Exp	-0.35 (-0.32,-0.37)*	Time at Min Exp	37%	54%	9%	-3 (-4, -3)*
			AUC	-0.42 (0.18,-1.02)	Time at Max+Min Exp	.	9%	.	
	9/13	234	Min Exp	-0.91 (-0.81,-1.01)*	Time at Min Exp	36%	42%	21%	-4 (-18, 4)*
			AUC	-17.2 (-14.8,-19.5)*	Time at Max+Min Exp	.	4%	.	
	11/13	423	Min Exp	-0.93 (-0.88,-0.99)*	Time at Min Exp	43%	41%	17%	-3 (-4, 2)*
			AUC	-12.5 (-11.1,-13.9)*	Time at Max+Min Exp	.	5%	.	

[1] Median (IQR) represents the median ignoring zero-differences.
*p-value $< 1.1 \times 10^{-4}$ (Bonferroni adjustment for 455 total tests conducted); Differences in continuous and ordinal features are tested with a one-sample t-test and Wilcoxon signed rank test, respectively.

Fourteen groups of genes were found to be clustered into different clusters, both with an overall initial decreasing trend (Figure 4.5). Table 4.5 summarizes comparisons between minimum and time to minimum expression. No significant differences in time to minimum expression were observed in AJ/B6 cluster 10/9, 10/11, 10/15, 10/12, 14/12, 10/13, 11/10, and 11/18, but a significant difference in the magnitude of minimum expression was observed.

AJ mice had a lower magnitude of minimum expression than B6 mice for genes in 10/9, 10/11, 10/15, 10/12, and 14/12, Conversely, B6 mice had a lower magnitude of minimum expression than AJ mice for genes in 10/13, 11/10, and 11/18.

Genes in 11/15 showed a significant horizontal shift to an later time at minimum in AJ mice versus B6 mice, but no differences in the magnitude of minimum expression was observed. Finally, genes in 17/15, 9/10, 17/18, 9/13 and 11/13 all showed significant horizontal and vertical shifts. All but 9/10 showed B6 having a shorter time to minimum expression. All groups except for 17/18 showed significant differences in AUC between strains.

Enriched KEGG pathways for genes that showed longer time to maximum or minimum expression for AJ mice are listed in Supplementary Table B.9 and B.10, respectively.

4.4.4 AJ versus C3H comparison

We can conduct similar comparisons between AJ and C3H. Table 4.6 is a cross-tabulation of the cluster labels for temporal gene-expression profiles belonging to strain AJ versus C3H. There were 3 groups of genes (Supplementary Figure B.5) that were found such that one strain showed an initial increase of expression compared to the other strain with an initial decrease in expression. Additionally, there were 4 and 7 groups of genes where one strain showed an initial increase or decrease in expression compared to a relatively flat trend in the other strain (Supplementary Figure B.6 and B.7, respectively).

Eight groups of genes were found to be clustered into different clusters both with an overall initial increasing trend (Figure 4.6). Table 4.7 summarizes the differences in features between AJ and C3H strain grouped by type of difference. No significant strain-specific differences in maximum expression or time to maximum expression among genes in AJ/C3H cluster 2/3.

Similarly, no significant differences in maximum expression was observed in

Table 4.6: Cross-tabulation of AJ versus C3H cluster labels. The shaded cells on-diagonal correspond to genes that were grouped into the same cluster for the AJ and B6 strain.

AJ Strain \ C3H Strain	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
1	1	0	4	17	5	0	0	0	0	0	0	0	0	0	0	0	0	0
2	7	1	109	493	25	2	1	0	0	0	0	0	0	0	0	0	0	0
3	12	1	430	128	156	145	9	3	4	2	0	1	0	0	0	0	0	0
4	20	1	66	619	347	19	3	5	0	0	0	0	0	0	0	0	0	0
5	8	0	34	183	1000	58	325	112	12	16	1	5	0	0	0	0	0	0
6	0	0	81	10	156	413	83	65	106	26	2	2	0	0	0	0	0	0
7	1	0	2	2	236	33	554	417	58	172	2	32	16	1	1	1	0	0
8	0	0	2	5	191	40	292	400	105	204	4	23	3	0	0	0	0	0
9	0	0	4	1	62	102	153	488	405	475	46	24	9	3	2	7	0	0
10	0	0	1	0	106	35	506	2278	384	2407	88	156	##	62	26	40	7	26
11	0	0	1	0	0	0	14	35	117	409	120	46	49	119	19	267	17	71
12	0	0	0	0	1	0	25	8	1	84	2	64	34	33	3	2	0	16
13	0	0	0	0	1	0	17	33	4	150	13	11	24	14	4	12	2	10
14	0	0	0	0	0	0	3	1	1	41	2	32	43	78	2	35	2	372
15	0	0	0	0	0	0	3	4	1	135	30	23	36	101	25	104	15	54
16	0	0	0	0	0	1	0	23	10	141	31	6	8	21	17	28	7	47
17	0	0	0	0	0	0	0	0	1	71	10	12	30	133	14	724	28	602
18	0	0	0	0	0	1	0	0	0	16	3	1	1	9	2	15	8	34

AJ/C3H cluster 3/5, 6/5, 5/4, 3/4, 2/4, 3/6. However, AJ mice has a significantly longer time to maximum expression than C3H mice in all groups except for 3/6 (where genes from C3H showed a longer time to maximum expression). Lastly, AJ/C3H cluster 4/5 showed a significant horizontal and vertical time to maximum expression and maximum expression, resulting in an earlier time to maximum with a higher magnitude of maximum expression. The AUC was significantly different between strains for 2/3, 6/5 and 2/4.

Seventeen groups of genes were found to be clustered into different clusters, both with an overall initial decreasing trend (Figure 4.7). Similar to before, Table 4.8 summarizes comparisons between minimum and time to minimum expression. No significant differences in time to minimum expression were observed in AJ/B6 cluster 11/10, 15/16, 11/14, 17/14,10/13, 15/10, 14/18, 13/10, and 16/10, but a significant difference in the magnitude of minimum expression was observed. Of these sets, all except for 13/10 and 16/10, C3H mice had a lower magnitude of minimum expression than AJ mice for genes. Genes in 15/14, 10/12, and 10/9 showed a significant horizontal shift for time to minimum, but no differences in the

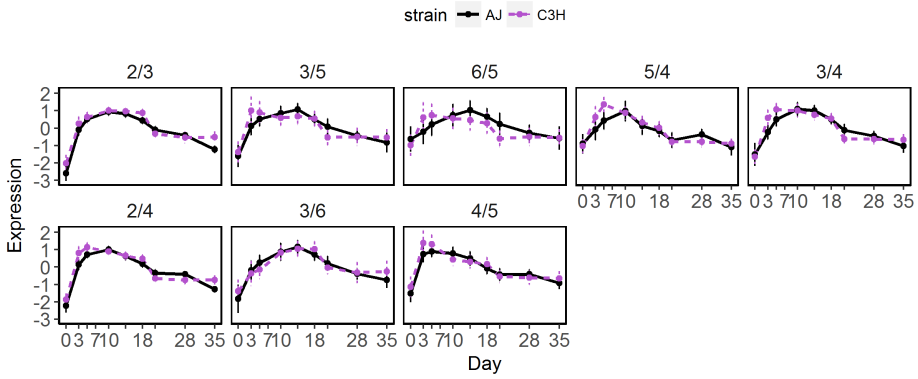

Figure 4.6: Strain-specific mean (SD) time-curve over all genes that were clustered into two different clusters across strain AJ and C3H, where both clusters showed an initial increasing trend

Table 4.7: Summary of features (area under the curve (AUC), maximum expression (Exp), and time at maximum Exp) comparing strain differences (C3H-AJ) in temporal gene expression curves clustered into groups with an initial increasing trend.

			Continuous Features		Ordinal Features				
	Cluster			Mean of		Sign of Diff			Median of
Type	AJ/C3H	N Genes	Parameter	C3H-AJ (95% CI)	Parameter	Negative	Zero	Positive	C3H-AJ (IQR)
No horizontal shift	2/3	109	Max Exp	0.05 (0.01, 0.09)	Time at Max Exp	19%	52%	28%	3 (-3, 7)
No vertical shift			AUC	4.66 (3.20, 6.13)*	Time at Max+Min Exp	.	34%	.	
Horizontal shift	3/5	156	Max Exp	0.06 (-0.04, 0.17)	Time at Max Exp	72%	18%	10%	-7 (-11, -2)*
No vertical shift			AUC	-2.39 (-4.97, 0.20)	Time at Max+Min Exp	.	8%	.	
	6/5	156	Max Exp	-0.04 (-0.16, 0.08)	Time at Max Exp	71%	20%	9%	-5 (-11, -3)*
			AUC	-7.70 (-11.1,-4.33)*	Time at Max+Min Exp	.	5%	.	
	5/4	183	Max Exp	-0.01 (-0.11, 0.08)	Time at Max Exp	53%	42%	5%	-2 (-3, -2)*
			AUC	1.27 (-0.44, 2.97)	Time at Max+Min Exp	.	22%	.	
	3/4	128	Max Exp	-0.06 (-0.13, 0.01)	Time at Max Exp	50%	41%	9%	-4 (-5, -2)*
			AUC	-0.93 (-2.61, 0.74)	Time at Max+Min Exp	.	16%	.	
	2/4	493	Max Exp	0.01 (-0.01, 0.02)	Time at Max Exp	50%	46%	4%	-2 (-4, -2)*
			AUC	1.56 (1.05, 2.07)*	Time at Max+Min Exp	.	25%	.	
	3/6	145	Max Exp	-0.06 (-0.15, 0.03)	Time at Max Exp	12%	47%	41%	4 (3, 4)*
			AUC	0.48 (-2.34, 3.29)	Time at Max+Min Exp	.	23%	.	
Horizontal shift	4/5	347	Max Exp	0.33 (0.27, 0.40)*	Time at Max Exp	68%	27%	5%	-4 (-4, -2)*
Vertical shift			AUC	0.30 (-0.99, 1.58)	Time at Max+Min Exp	.	11%	.	

[1] Median (IQR) represents the median ignoring zero-differences.

*p-value < 1.1×10^{-4} (Bonferroni adjustment for 595 total tests conducted). Differences in continuous and ordinal features are tested with a one-sample t-test and wilcoxon signed rank test, respectively.

magnitude of minimum expression was observed. Finally, genes in 11/9, 11/16, 17/16, 17/18, and 9/10 all showed significant horizontal and vertical shifts. All groups except for 13/10 showed significant differences in AUC between strains.

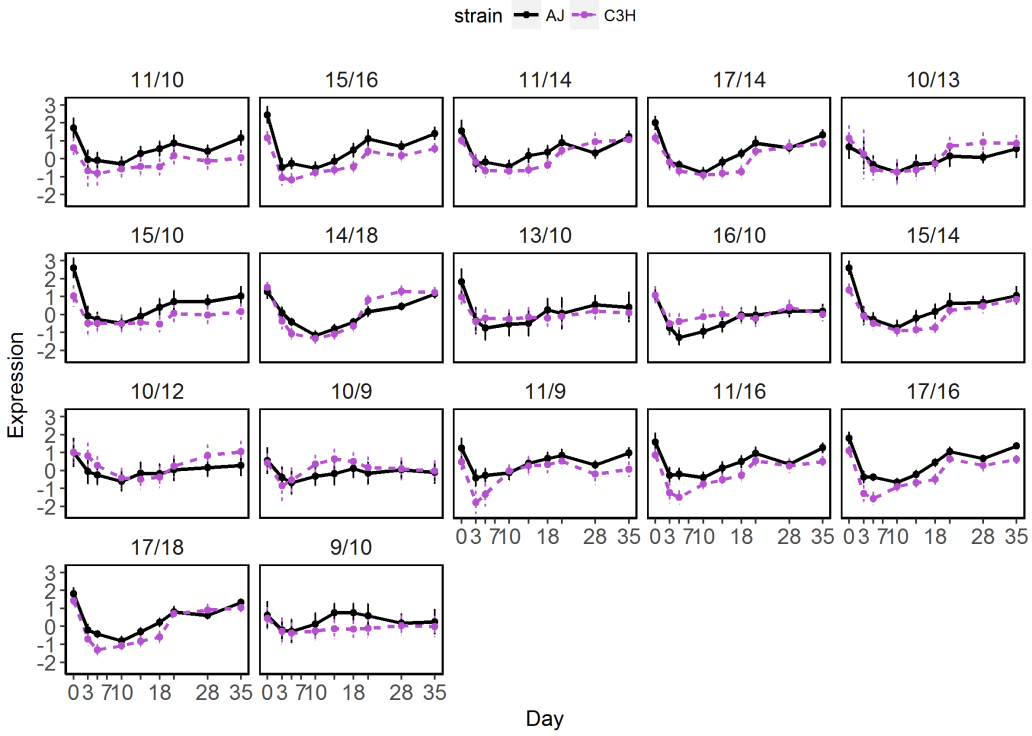

Figure 4.7: Strain-specific mean (SD) time-curve over all genes that were clustered into two different clusters across strain AJ and C3H, where both clusters showed an initial decreasing trend

Enriched KEGG pathways for genes that showed longer time to maximum or minimum expression for AJ mice are listed in Supplementary Table B.11 and B.12, respectively.

Table 4.8: Summary of features (area under the curve (AUC), minimum expression (Exp), and time at minimum Exp) comparing strain differences (C3H-AJ) in temporal gene expression curves clustered into groups with an initial decreasing trend.

			Continuous Features		Ordinal Features				
	Cluster			Mean of		Sign of Diff			Median of
Type	AJ/C3H	N Genes	Parameter	C3H-AJ (95% CI)	Parameter	Negative	Zero	Positive	C3H-AJ (IQR)
No horizontal shift	11/10	409	Min Exp	-0.72 (-0.78,-0.66)*	Time at Min Exp	44%	28%	29%	-2 (-4, 4)
Vertical Shift			AUC	-24.4 (-25.8,-23.0)*	Time at Max+Min Exp	.	12%	.	
	15/16	104	Min Exp	-0.48 (-0.56,-0.40)*	Time at Min Exp	49%	30%	21%	-2 (-4, 2)
			AUC	-22.1 (-23.8,-20.3)*	Time at Max+Min Exp	.	19%	.	
	11/14	119	Min Exp	-0.40 (-0.48,-0.33)*	Time at Min Exp	37%	23%	40%	2 (-3, 4)
			AUC	-7.65 (-9.88,-5.42)*	Time at Max+Min Exp	.	4%	.	
	17/14	133	Min Exp	-0.30 (-0.35,-0.24)*	Time at Min Exp	38%	44%	17%	-3 (-3, 3)
			AUC	-12.7 (-14.3,-11.2)*	Time at Max+Min Exp	.	19%	.	
	10/13	103	Min Exp	-0.30 (-0.42,-0.17)*	Time at Min Exp	40%	43%	17%	-3 (-4, 3)
			AUC	7.97 (4.99,10.95)*	Time at Max+Min Exp	.	20%	.	
	15/10	135	Min Exp	-0.24 (-0.33,-0.15)*	Time at Min Exp	34%	28%	38%	2 (-4, 7)
			AUC	-20.6 (-23.1,-18.0)*	Time at Max+Min Exp	.	13%	.	
	14/18	372	Min Exp	-0.22 (-0.25,-0.19)*	Time at Min Exp	31%	52%	17%	-3 (-4, 3)
			AUC	4.30 (3.39, 5.21)*	Time at Max+Min Exp	.	36%	.	
	13/10	150	Min Exp	0.49 (0.38, 0.61)*	Time at Min Exp	29%	33%	37%	3 (-4, 11)
			AUC	-2.95 (-5.93, 0.02)	Time at Max+Min Exp	.	11%	.	
	16/10	141	Min Exp	0.68 (0.60, 0.75)*	Time at Min Exp	49%	21%	30%	-2 (-4, 5)
			AUC	8.95 (7.31,10.58)*	Time at Max+Min Exp	.	9%	.	
Horizontal shift	15/14	101	Min Exp	-0.18 (-0.27,-0.09)	Time at Min Exp	19%	33%	49%	4 (-3, 7)*
No Vertical Shift			AUC	-13.2 (-15.9,-10.5)*	Time at Max+Min Exp	.	18%	.	
	10/12	156	Min Exp	0.16 (0.04, 0.28)	Time at Min Exp	21%	28%	51%	4 (-3, 8)*
			AUC	10.34 (6.94,13.75)*	Time at Max+Min Exp	.	8%	.	
	10/9	384	Min Exp	-0.05 (-0.16, 0.06)	Time at Min Exp	55%	22%	22%	-2 (-7, 2)*
			AUC	9.12 (6.87,11.37)*	Time at Max+Min Exp	.	4%	.	
Horizontal shift	11/9	117	Min Exp	-1.24 (-1.39,-1.09)*	Time at Min Exp	51%	40%	9%	-4 (-4, -2)*
Vertical shift			AUC	-16.7 (-19.4,-14.1)*	Time at Max+Min Exp	.	23%	.	
	11/16	267	Min Exp	-0.97 (-1.03,-0.92)*	Time at Min Exp	59%	27%	14%	-3 (-5, -2)*
			AUC	-19.3 (-20.5,-18.1)*	Time at Max+Min Exp	.	15%	.	
	17/16	724	Min Exp	-0.73 (-0.76,-0.70)*	Time at Min Exp	77%	19%	4%	-2 (-4, -2)*
			AUC	-20.2 (-20.7,-19.7)*	Time at Max+Min Exp	.	15%	.	
	17/18	602	Min Exp	-0.40 (-0.42,-0.37)*	Time at Min Exp	50%	46%	5%	-2 (-3, -2)*
			AUC	-9.83 (-10.6,-9.11)*	Time at Max+Min Exp	.	24%	.	
	9/10	475	Min Exp	-0.26 (-0.32,-0.20)*	Time at Min Exp	31%	21%	49%	2 (-3, 7)*
			AUC	-15.6 (-17.1,-14.1)*	Time at Max+Min Exp	.	4%	.	

[1] Median (IQR) represents the median ignoring zero-differences.

*p-value < 1.1×10^{-4} (Bonferroni adjustment for 595 total tests conducted). Differences in continuous and ordinal features are tested with a one-sample t-test and wilcoxon signed rank test, respectively.

Table 4.9: Cross-tabulation of B6 versus C3H cluster labels. The shaded cells on-diagonal correspond to genes that were grouped into the same cluster for the B6 and C3H strain of mouse.

B6 Strain \ C3H Strain	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
1	13	2	273	772	173	21	9	6	0	1	0	1	0	0	0	0	0	0
2	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	4	0	156	16	53	62	8	7	3	2	0	0	0	0	0	0	0	0
4	4	0	11	378	273	2	9	3	0	0	0	0	0	0	0	0	0	0
5	21	1	189	253	1118	139	461	268	36	37	1	12	0	0	0	0	0	0
6	0	0	62	4	111	260	69	214	79	52	8	2	0	0	2	0	0	0
7	4	0	23	30	338	157	621	1236	144	620	6	77	17	12	0	1	0	3
8	2	0	4	3	89	21	183	456	64	238	6	9	1	0	2	0	0	0
9	0	0	3	0	36	79	148	727	209	615	26	18	8	9	2	10	0	6
10	0	0	13	1	77	89	268	798	365	1418	64	90	59	44	37	21	5	29
11	0	0	0	0	0	0	9	20	18	141	28	9	24	31	9	23	6	17
12	0	0	0	0	4	0	104	36	9	206	4	100	59	70	6	12	1	75
13	0	0	0	0	13	18	77	78	240	556	107	70	57	99	7	105	6	38
14	0	0	0	0	1	1	7	4	4	48	4	14	44	49	11	20	3	302
15	0	0	0	0	0	0	12	13	8	216	32	17	28	89	16	231	13	103
16	0	0	0	0	0	0	1	4	26	57	27	6	14	7	9	56	11	23
17	0	0	0	0	0	0	1	0	1	31	6	2	11	29	3	179	8	176
18	0	0	0	0	0	0	1	2	3	111	35	11	34	135	11	577	33	460

4.4.5 B6 versus C3H comparison

Table 4.9 is a cross-tabulation of the cluster labels for temporal gene-expression profiles belonging to strain B6 versus C3H. There were no groups of genes of substantial size (N>100) that were found such that one strain showed an initial increase of expression compared to the other strain with an initial decrease in expression. Additionally, there were 5 and 8 groups of genes where one strain showed an initial increase or decrease in expression compared to a relatively flat trend in the other strain (Supplementary Figure B.8 and B.9, respectively). Eight groups of genes were found to be clustered into different clusters both with an overall initial increasing trend (Figure 4.8). Table 4.10 summarizes the differences in features between B6 and C3H strain grouped by type of difference. No significant differences in maximum expression was observed in B6/C3H cluster 6/5. However, B6 mice had a significantly earlier time to maximum expression than C3H mice. Significant differences in the magnitude of maximum expression for clusters 5/3 and 1/3 are observed (C3H showed a higher magnitude of maximum expression), however,

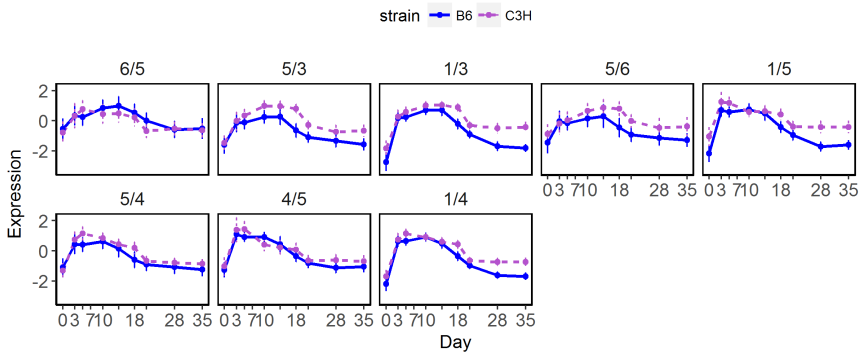

Figure 4.8: Strain-specific mean (SD) time-curve over all genes that were clustered into two different clusters across strain B6 and C3H, where both clusters showed an initial increasing trend.

Table 4.10: Summary of features (area under the curve (AUC), maximum expression (Exp), and time at maximum Exp) comparing strain differences (C3H-B6) in temporal gene expression curves clustered into groups with an initial increasing trend.

			Continuous Features		Ordinal Features				
	Cluster			Mean of		Sign of Diff			Median of
Type	B6/C3H	N Genes	Parameter	C3H-B6 (95% CI)	Parameter	-	0	+	C3H-B6 (IQR)
Horizontal shift	6/5	111	Max Exp	-0.11 (-0.27, 0.04)	Time at Max Exp	68%	19%	14%	-5 (-9, -2)*
No vertical shift			AUC	-7.78 (-12.3,-3.21)	Time at Max+Min Exp	.	7%	.	
No horizontal shift	5/3	189	Max Exp	0.71 (0.64, 0.79)*	Time at Max Exp	28%	38%	34%	3 (-4, 4)
Vertical shift			AUC	25.26 (23.25,27.27)*	Time at Max+Min Exp	.	19%	.	
	1/3	273	Max Exp	0.30 (0.26, 0.34)*	Time at Max Exp	27%	48%	26%	-2 (-4, 4)
			AUC	25.36 (24.09,26.63)*	Time at Max+Min Exp	.	32%	.	
Horizontal shift	5/6	139	Max Exp	0.56 (0.43, 0.69)*	Time at Max Exp	19%	28%	53%	4 (-2, 7)*
Vertical shift			AUC	22.69 (19.31,26.07)*	Time at Max+Min Exp	.	6%	.	
	1/5	173	Max Exp	0.40 (0.32, 0.48)*	Time at Max Exp	65%	28%	7%	-4 (-4, -2)*
			AUC	23.03 (20.98,25.08)*	Time at Max+Min Exp	.	15%	.	
	5/4	253	Max Exp	0.39 (0.32, 0.46)*	Time at Max Exp	40%	44%	17%	-2 (-3, 2)*
			AUC	12.04 (10.23,13.86)*	Time at Max+Min Exp	.	21%	.	
	4/5	273	Max Exp	0.27 (0.20, 0.35)*	Time at Max Exp	59%	30%	11%	-2 (-4, -2)*
			AUC	5.68 (4.05, 7.31)*	Time at Max+Min Exp	.	14%	.	
	1/4	772	Max Exp	0.17 (0.15, 0.19)*	Time at Max Exp	48%	43%	9%	-2 (-3, -2)*
			AUC	16.68 (16.08,17.28)*	Time at Max+Min Exp	.	29%	.	

[1] Median (IQR) represents the median ignoring zero-differences.

*p-value $< 1.1 \times 10^{-4}$ (Bonferroni adjustment for 595 total tests conducted). Differences in continuous and ordinal features are tested with a one-sample t-test and wilcoxon signed rank test, respectively.

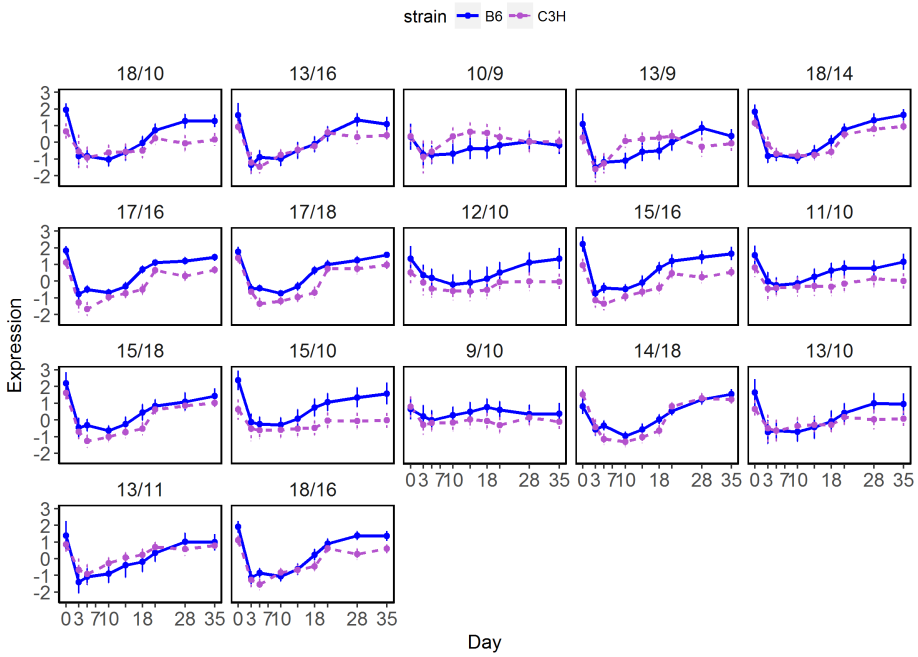

Figure 4.9: Strain-specific mean (SD) time-curve over all genes that were clustered into two different clusters across strain B6 and C3H, where both clusters showed an initial decreasing trend.

time to maximum for both strains were not different. Clusters 5/6, 1/5, 5/4, 4/5 and 1/4 showed significant differences in the magnitude and time of maximum expression, where C3H mice had a higher magnitude of expression. All groups except for 5/6 showed an earlier time to maximum for C3H mice. All groups except for 6/5 showed significant differences in AUC between strains.

Seventeen groups of genes were found to be clustered into different clusters, both with an overall initial decreasing trend (Figure 4.9). Table 4.11 summarizes comparisons between minimum and time to minimum expression. No significant

Table 4.11: Summary of features (area under the curve (AUC), maximum expression (Exp), and time at maximum Exp) comparing strain differences (C3H-B6) in temporal gene expression curves clustered into groups with an initial decreasing trend.

			Continuous Features		Ordinal Features				
	Cluster			Mean of		Sign of Diff			Median of
Type	B6/C3H	N Genes	Parameter	C3H-B6 (95% CI)	Parameter	-	0	+	C3H-B6 (IQR)
No Horizontal shift	18/10	111	Min Exp	0.01 (-0.09, 0.11)	Time at Min Exp	41%	23%	35%	-2 (-4, 3)
No vertical shift			AUC	-16.1 (-18.7,-13.5)*	Time at Max+Min Exp	.	5%	.	
	13/16	105	Min Exp	-0.07 (-0.15, 0.01)	Time at Min Exp	42%	22%	36%	-2 (-4, 2)
			AUC	-9.50 (-11.7,-7.29)*	Time at Max+Min Exp	.	6%	.	
Horizontal shift	10/9	365	Min Exp	0.05 (-0.02, 0.13)	Time at Min Exp	58%	25%	17%	-4 (-7, -2)*
No vertical shift			AUC	16.12 (14.04,18.20)*	Time at Max+Min Exp	.	2%	.	
	13/9	240	Min Exp	-0.02 (-0.10, 0.06)	Time at Min Exp	44%	40%	16%	-4 (-4, 2)*
			AUC	3.80 (1.98, 5.61)*	Time at Max+Min Exp	.	2%	.	
	18/14	135	Min Exp	0.06 (-0.01, 0.12)	Time at Min Exp	19%	36%	45%	3 (-2, 7)*
			AUC	-9.10 (-11.2,-7.04)*	Time at Max+Min Exp	.	22%	.	
No horizontal shift	17/16	179	Min Exp	-0.72 (-0.77,-0.67)*	Time at Min Exp	47%	27%	26%	-2 (-2, 2)
Vertical shift			AUC	-24.0 (-25.0,-22.9)*	Time at Max+Min Exp	.	12%	.	
	17/18	176	Min Exp	-0.65 (-0.69,-0.60)*	Time at Min Exp	28%	61%	11%	-2 (-3, 2)
			AUC	-20.2 (-21.6,-18.8)*	Time at Max+Min Exp	.	27%	.	
	12/10	206	Min Exp	-0.60 (-0.70,-0.50)*	Time at Min Exp	45%	21%	34%	-2 (-7, 4)
			AUC	-26.8 (-30.0,-23.6)*	Time at Max+Min Exp	.	8%	.	
	15/16	231	Min Exp	-0.56 (-0.60,-0.51)*	Time at Min Exp	35%	29%	36%	2 (-2, 2)
			AUC	-29.9 (-31.2,-28.5)*	Time at Max+Min Exp	.	19%	.	
	11/10	141	Min Exp	-0.48 (-0.58,-0.39)*	Time at Min Exp	45%	20%	35%	-2 (-4, 5)
			AUC	-22.5 (-25.2,-19.8)*	Time at Max+Min Exp	.	5%	.	
	15/18	103	Min Exp	-0.47 (-0.54,-0.39)*	Time at Min Exp	30%	35%	35%	2 (-2, 3)
			AUC	-15.3 (-17.7,-13.0)*	Time at Max+Min Exp	.	20%	.	
	15/10	216	Min Exp	-0.46 (-0.52,-0.39)*	Time at Min Exp	34%	27%	39%	2 (-3, 5)
			AUC	-33.8 (-36.4,-31.2)*	Time at Max+Min Exp	.	6%	.	
	9/10	615	Min Exp	-0.43 (-0.48,-0.38)*	Time at Min Exp	39%	15%	46%	2 (-4, 9)
			AUC	-16.8 (-18.2,-15.5)*	Time at Max+Min Exp	.	3%	.	
	14/18	302	Min Exp	-0.37 (-0.41,-0.33)*	Time at Min Exp	28%	44%	28%	-2 (-3, 3)
			AUC	-5.54 (-6.78,-4.29)*	Time at Max+Min Exp	.	42%	.	
	13/10	556	Min Exp	0.10 (0.05, 0.15)*	Time at Min Exp	43%	21%	36%	-2 (-4, 4)
			AUC	-10.3 (-11.7,-8.89)*	Time at Max+Min Exp	.	5%	.	
	13/11	107	Min Exp	0.48 (0.37, 0.59)*	Time at Min Exp	29%	31%	40%	2 (-2, 2)
			AUC	6.01 (3.13, 8.90)*	Time at Max+Min Exp	.	12%	.	
Horizontal shift	18/16	577	Min Exp	-0.24 (-0.26,-0.21)*	Time at Min Exp	59%	19%	22%	-2 (-3, 2)*
Vertical shift			AUC	-16.1 (-16.8,-15.3)*	Time at Max+Min Exp	.	6%	.	

[1] Median (IQR) represents the median ignoring zero-differences.

*p-value < 1.1×10^{-4} (Bonferroni adjustment for 595 total tests conducted). Differences in continuous and ordinal features are tested with a one-sample t-test and wilcoxon signed rank test, respectively.

differences in the magnitude of minimum or time to minimum expression were observed in B6/C3H cluster 18/10 and 13/16. A horizontal shift in time to minimum was observed in 10/9, 13/9 and 18/14, but no differences in the magnitude of minimum expression were seen. Groups 10/9 and 13/9 consist of genes that showed an earlier time to minimum for C3H. 18/14 showed a longer time to minimum for C3H strain. Genes in 17/16, 17/18, 12/10, 15/16, 11/10, 15/18, 15/10, 9/10, 14/18, 13/10 and 13/11 showed genes with a significant difference in magnitude of minimum expression, but not time of minimum expression. In all groups except for 13/10 and 13/11, B6 had a higher magnitude of minimum expression compared to C3H strain. Lastly, 18/16 consist of genes with a higher magnitude of minimum expression for B6 with an earlier time to minimum. All groups showed significant differences in AUC between strains.

Enriched KEGG pathways for genes that showed longer time to maximum or minimum expression for B6 mice are listed in Supplementary Table B.13 and B.14, respectively.

4.4.6 Same pattern over all 3 strains

From these results, we can also determine the genes that did not change clusters across all three strains, which could consist of genes not affected by genetic variation during the fracture healing process. Figure 4.10 shows these groups of genes. There are no visual differences in the temporal profile for the genes within each panel, suggesting that the clustering algorithm correctly grouped them together.

The types of signal transduction or regulatory pathways associated with these genes were found by determining defined KEGG pathways found to be statistically associated with them. (Table 4.12)

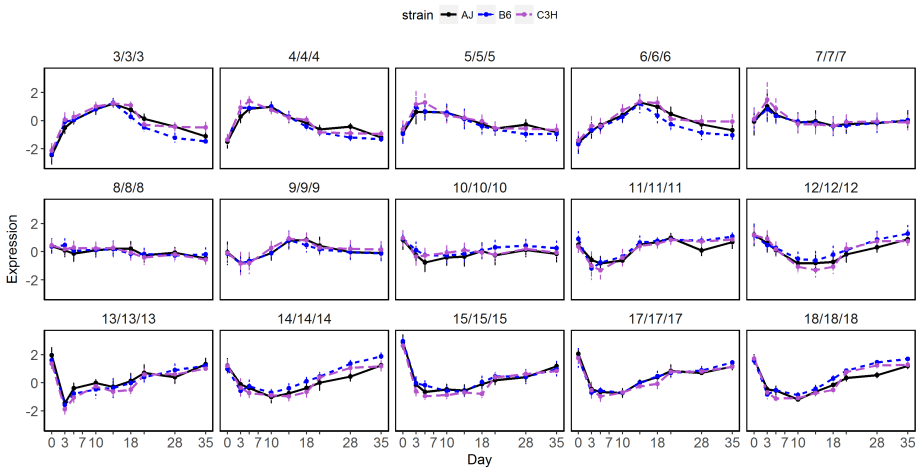

Figure 4.10: Gene expression profiles for genes that were clustered into the same cluster across all three strains (AJ, B6 and C3H).

Table 4.12: Identified KEGG[1] pathways in the genes grouped into the same cluster across all three strains.

Pathway Name	**No. Genes**
Olfactory transduction	176
Hypertrophic cardiomyopathy (HCM)	26
Arrhythmogenic right ventricular cardiomyopathy (ARVC)	23
Dilated cardiomyopathy (DCM)	25
TGF-beta signaling pathway	24
Lysosome	28
Hippo signaling pathway	33
Proteoglycans in cancer	39
Collecting duct acid secretion	9
Metabolic pathways	172
Melanogenesis	22
Amino sugar and nucleotide sugar metabolism	13
Primary bile acid biosynthesis	6
Adrenergic signaling in cardiomyocytes	29
Synaptic vesicle cycle	15

[1](Kyoto Encyclopedia of Genes and Genomes (KEGG) Pathways: (http://www.genome.jp/kegg/pathway.html)

4.5 DISCUSSION

In this chapter, we applied the modified initialization EPEM clustering algorithm (I-EPEM) to the bone fracture healing data to determine strain-specific differences in temporal gene-expression patterns. The I-EPEM algorithm, developed in the previous chapter, was used to cluster the data by treating expression profiles for each strain as separate objects to cluster. From simulation studies, we found that if large enough differences existed between factors, the clustering algorithm would successfully partition the genes into different clusters. However, when the differences were small and within-replicate variability was high ($\sigma = 0.4$), the clustering process tended to group them together. The results were what we expected, as the purpose of any cluster analysis is to group like elements together and different elements apart, even despite the added correlation introduced to the clustering (multiple curves for the same gene). However, a more extensive simulation study may be needed to determine if even higher levels of correlation (i.e. 0.99) or correlation introduced across all time-points may result in poor performance of I-EPEM.

In the bone fracture-healing data, several distinct temporal gene-expression patterns were identified for the different strains. Sets of genes were determined that exhibited different types of strain-specific differences: (1) no temporal trend for for either strain, (2) no temporal trend versus an initial increasing temporal trend, (3) no temporal trend versus an initial decreasing temporal trend, (4) both increasing temporal trends grouped into different clusters, (5) both decreasing temporal trends grouped into different clusters, and (6) one increasing and one decreasing temporal trend. For types 1, 2, 3, and 6 the clustering was able to identify which sets of genes exhibited these patterns. For types 4 and 5, we used specific features of the curve to determine if a substantial vertical or horizontal shift oc-

curred despite the similar overall increasing or decreasing trend. Sets of genes with a shift in pattern of temporal gene-expression is important to distinguish because they could be directly related to the mechanism by which fracture healing differs between the different strains of mice found in prior studies (Jepsen et al., 2008).

Several sets of genes were identified with these shifts, and using KEGG pathways several pathways were enriched for these genes with shifted temporal gene-expression patterns. The next step is to conduct a more comprehensive pathway analysis to determine how these genes are specifically related to the fracture healing process. For example, Jepsen et. al found that the slower healing mice C3H, had an earlier induction of osteogenesis compared to AJ mice. From this analysis we found several sets of genes that showed an earlier time at minimum or maximum expression for C3H mice, which could be directly relevant to this difference in induction time. In fact, two signaling pathways from Supplementary Table B.11, Rap1 and Wnt signaling pathways, have been found to be a critical component in osteoclast differentiation (Zou et al., 2013) (Glass et al., 2005). Additionally, compared to B6 mice, genes with a longer time to maximum expression in B6 mice compared to C3H mice were enriched in the MAPK signaling pathway, which has been found to be to involved in the regulation of master transcription factors that controls the functions of chondrocytes, osteoblasts and osteoclasts (Thouverey & Caverzasio, 2015). Lastly, this analysis was also able to identify genes that exhibited the same temporal pattern over all strains, which is an important part of understanding the biology behind fracture healing.

One benefit to EPEM that we did not explore is that we are able to obtain cluster membership probabilities for each gene. Using these membership probabilities,

genes on the boundary of the cluster with lower membership probability, could be discarded. Fine-tuned clusters could be obtained given some membership probability threshold for inclusion. The flexibility of this approach is a bonus to this work and further work can be done to obtain a more homogeneous group of genes for pathway analysis. We also ran the clustering using the original EPEM algorithm and it took 335 hours to converge compared to 42 hours from I-EPEM. A non-trivial amount of time was saved using the modified approach.

In this chapter, we have shown the benefits of using I-EPEM to obtain groups of genes with different patterns in gene expression across different strains of mice. The results are supportive of the observations from previous studies of strain-specific differences in rates of bone fracture-healing (Jepsen et al., 2008) (Grimes et al., 2011). Being able to find these patterns in large genomic datasets is an extremely important step to fully understanding the complex biologic process related to bone fracture-healing.

CHAPTER 5
Conclusions

5.1 OVERVIEW

The goal of any cluster analysis is to find structure in an unlabeled dataset by organizing the data into homogeneous groups based on a measure of similarity. Data in the same cluster are more similar, and data in different clusters are more dissimilar. Cluster analysis is particularly important in big datasets, such as microarray gene expression data, where the abundance of information makes processing of that information difficult. When the data is further complicated by introducing a time-component to the expression measurements for each gene or taking multiple biological replicates of the same time point, the clustering process needs to account for the additional variability or correlation introduced into the data. Current clustering methods either do not account for this variability or require multiple steps to the process making it too computationally inefficient for large gene expression datasets. The purpose of this dissertation was to develop a one-step clustering algorithm that accounts for all levels of this variability while attempting to decrease the computational burden that many of the current methods currently suffer from.

5.2 ENTROPY PENALIZED EM CLUSTERING ALGORITHM

In **Chapter 2**, we proposed an extension to an entropy penalized EM clustering algorithm (EPEM) (Chamroukhi, 2015) by utilizing mixtures of mixed effects polynomial regression models and adding gene- and replicate-specific random-effect terms. The random-effects accounted for the variability in the data resulting from repeated measurements over time and replicates at each time point. Simulation

studies were conducted to determine how different mixed-effects model performed with the highest accuracy over a variety of scenarios.

We found that the addition of random-effects into our mixture model decreased the misclassification error of the clustering algorithm compared to mixtures of fixed-effects models (Chamroukhi, 2015) and other methods (Yang et al., 2012) (Fraley et al., 2012) (Hartigan & Wong, 1979). The impact of the replicate-specific effect was minimal, as models with or without a replicate-specific random effect resulted in clusters with similar accuracy. However, in large datasets, despite the low misclassification error of the algorithm, the computational burden was still high. The clustering algorithm was run on the fracture healing data accounting for strain-specific variability and 22 distinct clusters were found.

5.3 MODIFICATIONS TO EPEM

In **Chapter 3**, to address the high computational burden of the EPEM algorithm, two alterations to the original algorithm were proposed. A split-clustering algorithm was used to divide the data into subgroups, and cluster analysis was run for each of the separate subgroups to obtain cluster labels. However, the success of this method is dependent on how the data are initially divided. We proposed using model-selection methods using BIC to determine the optimal order for each temporal gene expression profile, and to use the optimal order to divide the data into curves with the same polynomial order. However, adding an additional level of classification on top of the clustering analysis resulted in a decreased accuracy of predicted cluster labels. If EPEM is run in parallel for each of the splits, then a large reduction in computation time is observed (8 versus 71 hours). For extremely large "omics" data, this significant reduction in time may be worth a small decrease

in accuracy.

The computational burden of the EPEM algorithm resides in the first few iterations of the EM algorithm as the method initializes with each temporal gene expression profile as its own cluster. In each iteration, the clusters are merged, and at the final iteration, groups of similar gene expression profiles are obtained. The second method is aimed at decreasing this initial burden in the first few iterations by pre-grouping the temporal gene expression profiles based on their estimated polynomial regression coefficients. Therefore, each initialization of the EPEM will start with 5^{p+1} groups, where p corresponds to the fixed-effect order of the polynomial regression model used to define the curves. Through simulation studies, we found that the method worked very well in predicting the cluster labels with low misclassification error. However, in situations where the between-replicate variability was much higher than the within-replicate variability ($\sigma^2 < \tau^2$), ignoring the replicate-specific effect in the mixed effects model resulted in an increase in misclassification error. If the variability within a temporal gene expression curve is low, but the variability between replicates is high, averaging over the results at each time point can result in more heterogeneous curves within a cluster. The model cannot account for this additional variability, and as a result, groups them into separate clusters. I-EPEM was run on the fracture healing data accounting for strain-specific variability and 20 distinct cluster were found. The results were compared to those obtained from EPEM, and found to be similar. However, I-EPEM took less than one-third of the time it took to run the original EPEM algorithm (20 hours versus 71).

Additionally, singular value decomposition (SVD) of the data matrix proved to be useful to extract subspaces of the data matrix. The subspaces were then sub-

jected to model selection techniques with AIC, BIC and cross-validation to determine the optimal polynomial order. From a simulation study, we found that this process worked well to determine the order (p) to use in our mixed effects model. However, we found that under-specification of the fixed-effects order of the polynomial model lead to results with a higher misclassification error, whereas over-specification of the model had little impact on the misclassification error. Generally, it is much better to over-specify the model so that each of the curves can be sufficiently represented for the clustering.

5.4 STRAIN-SPECIFIC DIFFERENCES IN PATTERNS OF TEMPORAL EXPRESSION

In **Chapter 4**, we applied the modified initialization clustering algorithm (I-EPEM) to determine strain-specific differences in patterns of temporal gene-expression curves in the bone-fracture healing data. Treating gene-expression profiles from each strain as separate objects to cluster allowed us to determine sets of genes that exhibited strain-specific differences. A simulation study determined that the additional correlation introduced into the data (between repeated measurements of the same gene across strain) did not negatively affect the results if strain-specific differences were large.

The clustering algorithm was able to divide the data into gene sets with differing patterns: (1) no temporal train for both strains, (2) no temporal trend versus an initial increasing temporal trend, (3) no temporal trend versus an initial decreasing temporal trend, (4) both increasing temporal trends grouped into different clusters, (5) both decreasing temporal trends grouped into different clusters or (6) one increasing and one decreasing temporal trend.

For types 4 and 5, pairwise differences in expression profiles between sets of genes identified from I-EPEM were conducted using measures borrowed from traditional pharmacokinetics studies. Specific features of the curve (area under the curve, time to maximum/minimum expression or magnitude of maximum/minimum expression) were used to determine if a substantial vertical or horizontal shift occurred despite the similar overall increasing or decreasing trend. Identification of sets of genes with shifted temporal-expression profiles were important to determine as previous studies have found differences in fracture-healing rates between the different strains.

5.5 LIMITATIONS AND FUTURE WORK

Some limitations of this work is that we are assuming the data follow a multivariate Gaussian distribution. The accuracy of the predicted clusters would decrease if the data of interest is very highly non-normal. Future work could extend the estimation procedure to account for other distributions such as the Student's t-distribution (in the presence of outliers) or the Gamma distribution (for heavily skewed data).

Another limitation is that we are assuming the temporal gene expression curves to be well represented by polynomial functions. However, this may not be the case, and other studies have found that using splines have increased the accuracy of the clustering results (Chamroukhi, 2015). In this particular application, with only ten time points, spline modeling may not be appropriate with so few time points. However, in other applications, where many more time points are obtained, this may be a better approximation to the time curve.

When assessing strain-specific differences, comparing differences in time to

maximum or minimum using a Wilcoxon signed rank test may not be the best method. In this case, time is an ordinal variable and restricted to the time points obtained from the experiment. Many zero differences in time to maximum/minimum were observed due to the restricted samples obtained at the set number of time points. If the interest of a gene-expression experiment is to truly test for these differences, then more frequent measurements around suspected maximums or minimums would be necessary for the comparison to be more accurate.

Lastly, we know that not all genes are independent and many act in unison to serve one purpose. The cluster algorithm, however, assumes that each temporal gene profile is independent of one another. Therefore, an extension to account for correlation between genes might result in better defined clusters.

5.6 DISCUSSION

In this dissertation, we have proposed a clustering algorithm that accounts for correlation between repeated measurements over time and replicate measurements at each time point in temporal gene expression data. We found the method to outperform other methods in big datasets. Additionally, a modification to the algorithm proved to be beneficial when time is a factor and data size is very large. In applying our data to a study assessing differences in patterns of expression over time from different strains of mice, we were able to define several sets of genes that may be related to previously observed differences in rates of fracture healing.

This work can be extended to any cluster analysis where continuous measurements are obtained over multiple time-points and the interest is to identify individuals or objects with similar time profiles. For example, we could use health measurements collected over time to group patients into several clusters with sim-

ilar time-profiles. These clusters could be used to determine groups with a higher risk of disease or groups to recruit for a clinical trial. The results of this dissertation have showed the importance of cluster analysis in identifying patterns in the data that may represent observed biologic phenomena that prior research have not yet been able to define.

APPENDIX A

Appendix

A.1 BONE HEALING MICROARRAY DATA

A.1.1 Animals

Animal research was conducted in conformity with an IACUC approved protocol. In the fracture healing study, C57BL/6J (B6) C3H/HeJ (C3H) and Aj strains of mice were purchased from Jackson Laboratories (Bar Harbor, ME). All fracture studies were performed on 8 to 10 week old male mice as previously described (Jepsen et al., 2008). Total RNA samples were assessed in a three pooled samples from duplicate fractured calluses (N=3 mice per time point) harvested on days 0 (no fracture), days 3, 5, 7, 10, 14, 18, 21, 28 and 35.

A.1.2 Fracture Model

Through a surgical procedure, unilateral, transverse fractures were generated to study the progression of bone healing starting at day 0. Fracture calluses (the hard or soft callus that forms during bone regeneration) were obtained by harvesting the fractured femora from euthanized mice on post-operative days 3, 5, 7, 10, 14, 18, 21, 28 and 35 or 3, 6, 9, 12 and 18 months for the bone fracture and aging study, respectively.

A.1.3 Microarray Analysis

All procedures were performed at Boston University Microarray Resource Facility exactly as described in GeneChip© Whole Transcript (WT) Sense Target Labeling Assay Manual (Affymetrix, Santa Clara, CA, current version available at

www.affymetrix.com). Triplicate mRNA pools made from a randomized pooling of mRNAs isolated from N=6 callus were used. The GeneChip © Mouse Gene 1.0ST Arrays were used for these studies. 200ng of RNA from each of the three mRNA pools was labeled and used for hybridization. After hybridizations microarrays were immediately scanned using Affymetrix GeneArray Scanner 3000 7G Plus (Affymetrix, Santa Clara, CA). File output was in CEL format and was summarized using Affymetrix Expression Console (current version 1.1). RMA (Robust Multi-Array Analysis) algorithm (Irizarry et al., 2003) was used to generate gene-level data. Transcript-level expression values were derived using the RMA Console to identify outlier arrays and batch effects. Data quality was assessed by the relative log expression (RLE) values of all probe sets, the Normalized Unscaled Standard Error (NUSE) values, and the area under the receiver operating characteristics (ROC) curve comparing signal values for positive and negative control probes. The expression values were log-base 2 transformed and standardized by the overall gene-specific mean and standard deviation for each dataset. Each array consisted of expression values for 21,187 (bone fracture) for a mouse harvested at a specific time point.

A.2 THE ENTROPY-PENALIZED LOG-LIKELIHOOD FUNCTION

Assume that the data consist of N temporal gene-expression profiles. Each profile consists of gene-expression measurements obtained at T time points, represented by y_i where i = 1, . . . , N. Now assume that for each time-point we have R replicates so that each temporal gene-expression profile has T$\times$R measurements. Recall that each cluster is allowed to have its own set of parameters and probability density function. Therefore, the mixture density of y_i can be represented

as a weighted sum of each component of the mixture model (A.1), where $\Theta = \{\theta_1, \ldots, \theta_K, \alpha_1, \ldots, \alpha_K\}$ and $\theta_k = \{\beta_k, \sigma_k^2\}$ are the cluster-specific parameters for the k-th cluster ($k = 1, \ldots, K$).

$$p(y_i|\Theta) = \sum_{k=1}^{K} \alpha_k p_k(y_i|\theta_k) \tag{A.1}$$

Recall that z_i represents the missing class label for temporal profile for gene i, and $z_i = c(z_{i1}, \ldots, z_i)$ where $z_{ik} = \mathbb{I}\{y_i \text{originated from cluster k}\}$ and $\sum_{k=1}^{K} z_{ik} = 1$. The complete data vector is (y_i, z_i). By Bayes Rule, if we assume that gene i originated from cluster k, the joint probability density function (PDF) for one realization of the complete-data, $d_i = (y_i, z_i)$, is $p(d_i|\theta_k) = \alpha_k p_k(y_i|\theta_k)$, then the PDF unconditional of cluster membership is (A.3).

$$p(d_i|\Theta) = \prod_{k=1}^{K} [\alpha_k p_k(y_i|\theta_k)]^{z_{ik}} \tag{A.2}$$

It follows that the observed-data penalized log-likelihood can be represented by (A.2).

$$\begin{aligned} l_{(p)}(\Theta) &= log[p(y|\Theta)] \\ &= log \prod_{i=1}^{N} \sum_{k=1}^{K} \alpha_k p_k(y_i|\theta_k) \\ &= \sum_{i=1}^{N} log\{\sum_{k=1}^{K} \alpha_k p_k(y_i|\theta_k)\} \\ &+ \lambda \sum_{i=1}^{N} \sum_{k=1}^{K} \alpha_k log \alpha_k \end{aligned} \tag{A.3}$$

A.3 THE E-STEP

The E-step ("Expectation" step) of the EM algorithm takes the expectation of the complete-data log-likelihood function (A.3) and conditioning it on the observed data and the parameter space of the previous iteration. We want to maximize the objective function shown in (A.4).

$$\begin{aligned} Q(\Theta, \Theta^{(s)}) &= \mathbb{E}[logL_c(\Theta)|y_i, \Theta^{(s)}] \\ &= \sum_{i=1}^{N}\sum_{k=1}^{K} w_{ik}^{(s)} log\alpha_k \\ &+ \sum_{i=1}^{N}\sum_{k=1}^{K} w_{ik}^{(s)} \mathbb{E}[h(\theta_k|y_i, b_{ik}, c_{ik})|y_i, \Theta^{(s)}] \\ &+ \lambda \sum_{i=1}^{N}\sum_{k=1}^{K} \alpha_k log\alpha_k \end{aligned} \tag{A.4}$$

where

$$\begin{aligned} h(\theta_k|y_i, b_{ik}, c_{ik}) = &-\frac{TRlog\sigma_k^2}{2} \\ &- \frac{||y_i - X\beta_k - Ub_{ik} - Vc_{ik}||^2}{2\sigma_k^2} \\ &- \frac{log|G_k|}{2} - \frac{b_i' G_k^{-1} b_i}{2} - \frac{Rlog\tau_k^2}{2} - \frac{c_i' c_i}{2\tau_k^2} \end{aligned} \tag{A.5}$$

w_{ik} represents the posterior probability that the i-th gene belongs to the k-th component conditioning on the observed data and the random effects (A.6).

$$\begin{aligned}
\hat{w}_{ik}^{(s)} &= \mathbb{E}[z_{ik}|y_i, \Theta^{(s)}] \\
&= P(z_i = k|\Theta^{(s)}, y_i) \\
&= \frac{P(y_i|z_i = k, \Theta^{(s)})P(z_i = k)}{P(y_i|\Theta^{(s)})} \\
&= \frac{\alpha_k^{(s)} p_k(y_i|\theta_k^{(s)})}{\sum_{k=1}^{K} \alpha_k^{(s)} p_k(y_i|\theta_k^{(s)})}
\end{aligned} \tag{A.6}$$

Now, we need to find the expectation of each of the random effects, conditional on the observed data, $E[b_{ik}|y_i]$, $E[c_{ik}|y_i]$, $E[b_{ik}b_{ik}^{'}|y_i]$ and $E[c_{ik}^{'}c_{ik}|y_i]$. Using the properties of normal distributions, we can determine the conditional expectations.

$$\begin{aligned}
\mathbb{E}(b_{ik}|y_i, \Theta^{(s)}) &= G_k U^{'} \Gamma_k^{-1}(y_i - X\beta_k) \\
Cov(b_{ik}|y_i, \Theta^{(s)}) &= G_k - G_k U^{'} \Gamma_k^{-1} U G_k \\
\mathbb{E}(b_{ik}b_{ik}^{'}|y_i, \Theta^{(s)}) &= G_k - G_k U^{'} \Gamma_k^{-1} U G_k \\
&\quad + G_k U^{'} \Gamma_k^{-1}(y_i - X\beta_k)(y_i - X\beta_k)^{'} \\
&\quad \Gamma_k^{-1} U G_k
\end{aligned} \tag{A.7}$$

where $\Gamma_k^{-1} = (UG_kU^{'} + \tau_k^2 VV^{'} + \sigma_k^2 I_{TR})^{-1}$

$$\begin{aligned}
\mathbb{E}(c_{ik}|y_i, \Theta^{(s)}) &= \tau_k^2 V^{'} \Gamma_k^{-1}(y_i - X\beta_k) \\
Cov(c_{ik}|y_i, \Theta^{(s)}) &= \tau_k^2 I_R - \tau_k^4 V^{'} \Gamma_k^{-1} V \\
\mathbb{E}(c_{ik}^{'}c_i|y_i, \Theta^{(s)}) &= \tau_k^4 (y_i - X\beta_k)^{'} \Gamma_k^{-1} V V^{'} \Gamma_k^{-1} \\
&\quad (y_i - X\beta_k) + R\tau_k^2 - \tau_k^4 tr(V^{'} \Gamma_k^{-1} V)
\end{aligned} \tag{A.8}$$

Lastly, if $D_{ik} = y_i - X\beta_k - Ub_{ik} - Vc_{ik}$, then we need to find $\mathbb{E}(D_{ik}|y_i, \Theta^{(s)})$ and $\mathbb{E}(D_{ik}^{'}D_{ik}|y_i, \Theta^{(s)})$.

$$\begin{aligned}\mathbb{E}(D_{ik}|y_i, \Theta^{(s)}) &= y_i - X\beta_k - U\mathbb{E}(b_{ik}|y_i, \Theta^{(s)}) - V\mathbb{E}(c_{ik}|y_i, \Theta^{(s)}) \\ &= \sigma_k^2\Gamma_k^{-1}(y_i - X\beta_k)\end{aligned} \tag{A.9}$$

Similarly, we can find the following

$$\begin{aligned}Cov(D_{ik}|y_i, \Theta^{(s)}) &= UCov(b_i|y_i, \Theta^{(s)})U^{'} + VCov(c_i|y_i, \Theta^{(s)})V^{'} \\ &= U[G_k - G_kU^{'}\Gamma_k^{-1}UG_k]U^{'}V \\ &\quad [\tau_k^2 I_R - \tau_k^2 V^{'}\Gamma_k^{-1}V]V^{'} \\ \mathbb{E}(D_{ik}^{'}D_{ik}|y_i, \Theta^{(s)}) &= \sigma_k^2(y_i - X\beta_k)^{'}\Gamma_k^{-1}\Gamma_k^{-1}(y_i - X\beta_k) \\ &\quad + tr[U[G_k - G_kU^{'}\Gamma_k^{-1}UG_k]U^{'}V \\ &\quad [\tau_k^2 I_R - \tau_k^2 V^{'}\Gamma_k^{-1}V]V^{'}]\end{aligned} \tag{A.10}$$

A.4 THE M-STEP

The M-step involves maximizing (2.4) with respect to each of the parameters $(\alpha_k, \beta_k, \sigma_k^2, \tau_k^2, G_k)$. By taking the derivative with respect to β_k, σ_k^2, τ_k^2, G_k and α_k we can derive the formulas for the MLEs of each parameter of interest (to solve for α_k, we must use a Lagrange multiplier to satisfy the constraint that $\sum_{k=1}^{K}\alpha_k = 1$). Note that $(Y_i|X\beta_k, U, V) \sim N(X\beta_k, \Gamma_k)$, where $\Gamma_k = UGU^{'} + \tau_k^2 VV^{'} + \sigma_k^2 I_{TR}$

For G_k, using the properties of expectation and trace and taking the derivative with respect to G_k^{-1}, we can find the updating equation for G_k (A.11).

$$\frac{\partial Q(\Theta,\Theta^{(s)})}{\partial G_k^{-1}} = \sum_{i=1}^{N} w_{ik}^{(s)} \left\{ -\frac{G_k}{2} - \frac{\mathbb{E}(b_{ik}b_{ik}^{'}|y_i,\Theta^{(s)})}{2} \right\} = 0$$

$$\hat{G_k}^{(s+1)} = \frac{\sum_{i=1}^{N} w_{ik}^{(s)} \mathbb{E}(b_{ik}b_{ik}^{'}|y_i,\Theta^{(s)})}{\sum_{i=1}^{N} w_{ik}^{(s)}} \tag{A.11}$$

Taking the derivative with respect to τ_k^2, σ_k^2 and β_k (A.12), we obtain the respective updating equations (A.13).

$$\begin{aligned}
\frac{\partial Q(\Theta,\Theta^{(s)})}{\partial \tau_k^2} &= \sum_{i=1}^{N} w_{ik}^{(s)} \left\{ -\frac{R}{2\tau_k^2} - \frac{\mathbb{E}(c_{ik}^{'}c_{ik}|y_i,\Theta^{(s)})}{2\tau_k^4} \right\} = 0 \\
\frac{\partial Q(\Theta,\Theta^{(s)})}{\partial \sigma_k^2} &= \sum_{i=1}^{N} w_{ik}^{(s)} \left\{ -\frac{TR}{2\sigma_k^2} + \frac{\mathbb{E}(D_{ik}^{'}D_{ik}|y_i,\Theta^{(s)})}{2\sigma_k^4} \right\} = 0 \\
\frac{\partial Q(\Theta,\Theta^{(s)})}{\partial \beta_k} &= \frac{1}{2\sigma_k^2} \sum_{i=1}^{N} w_{ik}^{(s)} [-2X^{'}\mathbb{E}(D_{ik} + X\beta_k|y_i,\Theta^{(s)}) \\
&\quad + 2X^{'}X\beta_k] = 0
\end{aligned} \tag{A.12}$$

$$
\begin{aligned}
\hat{\tau_k}^{2(s+1)} &= \frac{\sum_{i=1}^{N} w_{ik}^{(s)} \mathbb{E}(c_{ik}^{'} c_{ik} | y_i, \Theta^{(s)})}{R \sum_{i=1}^{N} w_{ik}^{(s)}} \\
\hat{\sigma_k}^{2(s+1)} &= \frac{\sum_{i=1}^{N} w_{ik}^{(s)} \mathbb{E}(D_{ik}^{'} D_{ik} | y_i, \Theta^{(s)})}{TR \sum_{i=1}^{N} w_{ik}^{(s)}} \\
\hat{\beta_k}^{(s+1)} &= \frac{\sum_{i=1}^{N} w_{ik}^{(s)} (X^{'}X)^{-1} X^{'} \mathbb{E}(D_{ik} | y_i, \Theta^{(s)})}{\sum_{i=1}^{N} w_{ik}^{(s)}}
\end{aligned}
\tag{A.13}
$$

For α_k, we must take the derivative with respect to α_k and use the method of Lagrange multipliers (shown below) to obtain the updating equation (A.14).

$$
\begin{aligned}
\frac{\sum_{i=1}^{N} w_{ik}^{(s)}}{\alpha_k} + \lambda n log \alpha_k + \lambda n &= \eta \\
\sum_{k=1}^{K} \alpha_k \left\{ \frac{\sum_{i=1}^{N} w_{ik}^{(s)}}{\alpha_k} + \lambda n log \alpha_k + \lambda n \right\} &= \eta \sum_{k=1}^{K} \alpha_k \\
\sum_{k=1}^{K} \sum_{i=1}^{N} w_{ik}^{(s)} + \lambda n \sum_{k=1}^{K} \alpha_k log \alpha_k + \lambda n \sum_{k=1}^{K} \alpha_k &= \eta \sum_{k=1}^{K} \alpha_k \\
n + \lambda n \sum_{k=1}^{K} \alpha_k log \alpha_k + \lambda n &= \eta
\end{aligned}
$$

Plugging the formula for η back into our partial derivative and multiplying both

sides by α_k, the updating equation for α_k is obtained.

$$\hat{\alpha}_k^{(s+1)} = \frac{\sum_{i=1}^{N} w_{ik}^{(s)}}{N} + \lambda \hat{\alpha}_k^{(s)} \left\{ log\hat{\alpha}_k^{(s)} - \sum_{k=1}^{K} \hat{\alpha}_k^{(s)} log\hat{\alpha}_k^{(s)} \right\} \tag{A.14}$$

After plugging our conditional expectations into the formulas in (A.13) and (A.14), the updating equations are now derived (A.15).

$$\begin{aligned}
\hat{\beta}_k^{(s+1)} =& \frac{\sum_{i=1}^{N} w_{ik}^{(s)} [\hat{\beta}_k^{(s)} + \hat{\sigma}_k^{2(s)} (X'X)^{-1} X' \Gamma_k^{-1(s)} (y_i - X\hat{\beta}_k^{(s)})]}{W_i} \\
\hat{\tau}_k^{2(s+1)} =& \frac{1}{RW_i} \sum_{i=1}^{N} w_{ik}^{(s)} [\hat{\tau}_k^{4(s)} (y_i - X\hat{\beta}_k^{(s)})' \Gamma_k^{-1(s)} VV' \Gamma_k^{-1(s)} \\
& (y_i - X\hat{\beta}_k^{(s)}) + R\hat{\tau}_k^{2(s)} - \hat{\tau}_k^{4(s)} tr(\Gamma_k^{-1(s)} VV')] \\
\hat{G}_k^{(s+1)} =& \frac{1}{W_i} \sum_{i=1}^{N} w_{ik}^{(s)} [\hat{G}_k^{(s)} U' \Gamma^{-1(s)} (y_i - X\hat{\beta}_k^{(s)})(y_i - X\hat{\beta}_k^{(s)})' \\
& \Gamma_k^{-1(s)} U\hat{G}_k^{(s)} + \hat{G}_k^{(s)} - \hat{G}_k^{(s)} U' \Gamma_k^{-1(s)} UG_k^{(s)}] \\
\hat{\sigma}_k^{2(s+1)} =& \frac{1}{TRW_i} \sum_{i=1}^{N} w_{ik}^{(s)} [\hat{\sigma}_k^{4(s)} (y_i - X\hat{\beta}_k^{(s)})' \Gamma_k^{-1(s)} \Gamma_k^{-1(s)} \\
& (y_i - X\hat{\beta}_k^{(s)}) + tr(Cov(D_{ik}|y_i, \Theta^{(s)})] \\
\hat{\alpha}_k^{(s+1)} =& \frac{W_i}{N} + \lambda \hat{\alpha}_k^{(s)} \left\{ log\hat{\alpha}_k^{(s)} - \sum_{k=1}^{K} \hat{\alpha}_k^{(s)} log\hat{\alpha}_k^{(s)} \right\}
\end{aligned} \tag{A.15}$$

where $W_i = \sum_{i=1}^{N} w_{ik}^{(s)}$, $\Gamma_k^{-1} = (UG_kU' + \tau_k^2 VV' + \sigma_k^2 I_{TR})^{-1}$, and $Cov(D_{ik}|y_i, \Theta^{(s)}) = U[\hat{G}_k^{(s)} - \hat{G}_k^{(s)} U' \Gamma_k^{-1(s)} U\hat{G}_k^{(s)}]U' + V[\hat{\tau}_k^2 I_R - \hat{\tau}_k^{4(s)} V' \Gamma_k^{-1(s)} V]V'$.

The algorithm alternates between the E-step and the M-step until a convergence criteria is satisfied. In this case, we assess convergence from the estimated fixed-

effects parameters such that $\max_k \|\hat{\beta}_k^{(s+1)} - \hat{\beta}_k^{(s)}\| < \epsilon$. The final class labels are determined by finding (A.16) for $k \in \{1, ..., K\}$.

$$\hat{z}_i = \arg\max_k \hat{w}_{ik} \tag{A.16}$$

A.5 SINGULAR VALUE DECOMPOSITION

Let Y denote the N $\times$ RT gene-expression matrix (N=number of genes; R=number of replicates; T=number of time points), with rank v, where N $\geq$ RT and therefore v $\leq$ RT. Then y_{ij} is the expression level of the i-th gene in the j-th assay. Linking it back to the notation in A.1, y_i is the expression profile of the i-th gene, which consists of RT observations. Y can be decomposed into three matrices such that

$$Y = USV^{'} \tag{A.17}$$

where U is a N $\times$ RT matrix called the left singular vectors, S is a RT $\times$ RT diagonal matrix, and $V^{'}$ is a RT $\times$ RT matrix. The columns of U are called the left singular vectors, u_k, that forms an orthonormal basis for the assay expression profiles. The rows of $V^{'}$ are the right singular vectors of length RT, v_k, that forms an orthonormal basis for the gene expression profiles, which we call *eigenvectors*. S is a diagonal matrix of singular values, s_k. Each eigenvector explains a certain proportion of the variability, which can be determined by $s_k(\sum_{k=1}^{RT} s_i^2)^{-1}$.

We performed SVD by normalizing each gene's expression profile to have zero mean and unit standard deviation. The top L eigenvectors that explain >90% of the variability of the data are obtained and model selection is performed on these L eigenvectors (Section 2.1.1). In our simulation studies, we used L = 3.

A.6 ADJUSTED RAND INDEX

To compare a given clustering results to an external criteria, a measure of agreement is necessary. Given that each gene is assigned to only one cluster, measures of agreements between two different partitions can be defined.

Given a set of N objects S = $\{O_1, \ldots, O_N\}$, suppose U=$\{u_1, \ldots, u_N\}$ and

V=$\{v_1, \ldots, v_N\}$

represent two different partitions of the objects in S. Let U represent our external criteria and V is our clustering result. Let a be the number of pairs of ojects that are placed in the same cluster in U and in the same cluster in V. Let b be the number of pairs of objects in the same class in U, but not in the same cluster in V. Let c be the number of pairs of objects in the same class in V, but not in the same cluster in U. Let d be the number of pairs of objects in different clusters in both partitions. The quantities, a and d are number of agreements and b and c are the number of disagreements.

The Rand index (Rand, 1971) is equal to $\frac{a+d}{a+b+c+d}$, which lies between 0 and 1. When two partitions agree perfectly, the Rand index is 1. However, the expected value of the Rand index does not equal zero. Therefore, the adjusted Rand index was proposed (Hubert & Arabie, 1985), which assumes a generalized hypergeometric distribution to account for randomness of the cluster labels. Let n_{ij} is the number of objects in both clusters u_i and v_j, and n_i and n_j are the number of objects in class u_i and v_j, respectively, then the adjusted Rand index is defined as

$$\frac{\sum_{i,j}\binom{n_{ik}}{2} - \left\{\sum_i \binom{n_{i.}}{2}\sum_j \binom{n_{.j}}{2}\right\}/\binom{n}{2}}{\frac{1}{2}\left\{\sum_i \binom{n_{i.}}{2} + \sum_j \binom{n_{.j}}{2}\right\} - \left\{\sum_i \binom{n_{i.}}{2}\sum_j \binom{n_{.j}}{2}\right\}/\binom{n}{2}}] \tag{A.18}$$

APPENDIX B

Supplementary Tables and Figures

Table B.1: Parameters used to simulate data with 30 clusters with 200 genes per cluster. Replicate-specific variability (τ_k^2) was set to 0.01 for all genes. Correlation between random effects are set to be $\rho(b_0, b_1)$= -0.5, $\rho(b_0, b_2)$= 0.5 and $\rho(b_1, b_2)$=-0.7. Time-points = (0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.9, 1.0).

Cluster	β_{0k}	β_{1k}	β_{2k}	σ_k^2	Var(b_0)	Var(b_0)	Var(b_1)	Var(b_2)
1	0	1	0	0.1	0	0.15	0	
2	0.1	0.1	0	0.01	0.02	0.015	0	
3	-0.1	3	0	0.01	0.02	0.45	0	
4	0.5	0	0	0.05	0.1	0	0	
5	0.4	2	0	0.01	0.08	0.3	0	
6	0.6	-2	0	0.01	0.12	0.3	0	
7	1	-1	0	0.05	0.2	0.15	0	
8	0	-1	0	0.01	0	0.15	0	
9	2	-2	0	0.01	0.4	0.3	0	
10	3	-5	0	0.1	0.6	0.75	0	
11	-2	5	0	0.1	0.4	0.75	0	
12	8.5	-20	0	0.03	1.7	3	0	
13	-8.5	20	0	0.01	1.7	3	0	
14	-7	8	0	0.01	1.4	1.2	0	
15	1	-15.5	15	0.03	0.2	2.325	1.5	
16	-1	15.5	-25	0.03	0.2	2.325	2.5	
17	7	-20	10	0.01	1.4	3	1	
18	-4	25.5	-15	0.04	0.8	3.825	1.5	
19	0.1	-15	10	0.01	0.02	2.25	1	
20	-0.1	15	-10	0.01	0.02	2.25	1	
21	-0.1	15	-20	0.01	0.02	2.25	2	
22	-5	30	-30	0.01	1	4.5	3	
23	5	-30	30	0.01	1	4.5	3	
24	1.5	-17	12	0.03	0.3	2.55	1.2	
25	2.5	-19	17	0.01	0.5	2.85	1.7	
26	-2.5	-18	16	0.01	0.5	2.7	1.6	
27	-10	-15	30	0.01	2	2.25	3	
28	-3	38	-35	0.01	0.6	5.7	3.5	
29	3	-10	10	0.01	0.6	1.5	1	
30	-4	8	-7	0.01	0.8	1.2	0.7	

Table B.2: EPEM clustering results with different underlying models (NoM, Eq0r0 or Eq2r1 from Simulation A data with 8 clusters and 4 replicates and 200 or 500 genes (i.e. 50 or 125 genes per cluster). Average (SD) predicted cluster number ($\hat{K}$) and misclassification error (MCE %) over 1000 simulated datasets.

			50 Genes/Cluster		**125 Genes/Cluster**	
hline**Data**	**Error**[1]	**Model**	$\hat{K}$	**MCE%**	$\hat{K}$	**MCE%**
Sim A	Low	EPEM-Yang	8.8 (0.8)	5.1 (4.7)	14.8 (0.9)	36.3 (4.8)
		EPEM-Chamroukhi	8.9 (0.9)	5.9 (5.2)	15.1 (1.0)	38.0 (5.1)
		EPEM-Eq0r1 (p=2, q=0)	8.5 (0.6)	2.7 (3.4)	10.4 (1.0)	12.5 (5.1)
		EPEM-Eq1r1 (p=2, q=1)	8.3 (0.5)	1.5 (2.6)	8.6 (0.7)	2.9 (3.4)
		EPEM-Eq2r1 (p=2, q=2)	8.3 (0.5)	1.8 (2.9)	8.6 (0.7)	2.8 (3.3)
		EPEM-Eq2r0 (p=2, q=2)[2]	8.1 (0.5)	1.5 (3.7)	8.1 (0.3)	0.4 (1.7)
		KMeans	12.0 (0.19)	13.9 (1.38)	18.0 (0.00)	24.4 (1.92)
		GMM	8.3 (0.49)	1.1 (2.07)	9.4 (0.85)	6.0 (3.87)
Sim A	High	EPEM-Yang	9.1 (1.0)	7.1 (5.4)	15.2 (1.0)	38.6 (5.0)
		EPEM-Chamroukhi	9.0 (0.9)	2.8 (3.6)	15.0 (0.9)	38.0 (4.7)
		EPEM-Eq0r1 (p=2, q=0)	8.4 (0.6)	2.8 (3.6)	10.5 (1.0)	13.0 (5.2)
		EPEM-Eq1r1 (p=2, q=1)	8.3 (0.5)	1.6 (2.9)	8.6 (0.7)	3.0 (3.4)
		EPEM-Eq2r1 (p=2, q=2)	8.3 (0.5)	1.7 (2.9)	8.6 (0.7)	2.9 (3.4)
		EPEM-Eq2r0 (p=2, q=2)[2]	8.0 (0.5)	2.1 (4.3)	8.0 (0.3)	1.0 (3.0)
		KMeans	10.6 (1.83)	20.3 (5.46)	17.2 (2.90)	32.5 (7.15)
		GMM	8.3 (0.52)	1.6 (2.56)	9.4 (0.80)	7.2 (3.97)
Sim B	Mixed	EPEM-Yang	7.9 (1.1)	10.2 (10.5)	10.5 (1.2)	17.4 (6.4)
		EPEM-Chamroukhi	8.0 (0.6)	3.4 (5.8)	10.1 (1.1)	15.6 (5.7)
		EPEM-Eq0r1 (p=3, q=0)	7.8 (0.5)	3.1 (6.6)	10.0 (1.1)	12.5 (6.2)
		EPEM-Eq1r1 (p=3, q=1)	7.8 (0.5)	3.1 (6.6)	8.7 (0.8)	4.8 (4.4)
		EPEM-Eq2r1 (p=3, q=2)	7.7 (0.6)	3.7 (7.2)	8.3 (0.6)	1.9 (3.4)
		EPEM-Eq2r0 (p=3, q=2)[2]	7.8 (0.5)	2.2 (5.8)	8.1 (0.4)	1.0 (2.7)
		KMeans	12.5 (0.68)	21.1 (3.43)	13.2 (1.13)	23.7 (4.46)
		GMM	8.0 (0.13)	0.1 (0.87)	8.0 (0.08)	0.0 (0.34)

EPEM=Entropy Penalized EM Algorithm; FE: Fixed Effect; RE: Random Effect;
p: GMM: Standard Gaussian Mixture Model; FE polynomial order; q: RE polynomial order.
[1]Low: σ_k^2=0.01,τ_k^2=0.01; High: σ_k^2=0.1,τ_k^2=0.1; Mixed: $\sigma_k^2 \in (0.01 - 0.03)$; $\tau_k = 0.01$
[2]No Rep RE, c_{rik}.

Table B.3: Estimated parameters from clustering simulation A data (125 genes per cluster, 4 replicates, 4 clusters) with entropy-penalized EM algorithm used by Chamroukhi (fixed effects only). Only the 4 clusters closest in L2-Norm $|\hat{\beta} - \beta|$ to the true clusters are reported. Mean $\pm$ SD over 1000 iterations.

Parameter	Data	Cluster 1		Cluster 2		Cluster 3		Cluster 4	
Proportions	Error[1]	α_k	$\hat{\alpha_1}$	α_k	$\hat{\alpha_2}$	α_k	$\hat{\alpha_3}$	α_k	$\hat{\alpha_4}$
	(A) Low	0.25	0.1443 (0.0359)	0.25	0.1319 (0.0280)	0.25	0.1439 (0.0380)	0.25	0.1290 (0.0271)
	(A) High	0.25	0.1426 (0.0353)	0.25	0.1337 (0.0296)	0.25	0.1403 (0.0351)	0.25	0.1284 (0.0256)
	(B) Mixed	0.25	0.1387 (0.0396)	0.25	0.1438 (0.0436)	0.25	0.1406 (0.0417)	0.25	0.1428 (0.0372)
Fixed effects									
p=0		β_{01}	$\hat{\beta}_{01}$	β_{02}	$\hat{\beta}_{02}$	β_{03}	$\hat{\beta}_{03}$	β_{04}	$\hat{\beta}_{04}$
	(A) Low	0.5	0.5003 (0.0320)	0.1	0.1000 (0.0183)	0.5	0.4994 (0.0310)	-1	-0.9992 (0.0119)
	(A) High	0.5	0.4781 (0.0899)	0.1	0.1004 (0.0569)	0.5	0.5180 (0.0876)	-1	-1.0006 (0.0377)
	(B) Mixed	-1.05	-1.0512 (0.0201)	-2.3	-2.3090 (0.1634)	2.03	2.0416 (0.0889)	2.85	2.9081 (0.2869)
p=1		β_{11}	$\hat{\beta}_{11}$	β_{12}	$\hat{\beta}_{12}$	β_{13}	$\hat{\beta}_{13}$	β_{14}	$\hat{\beta}_{14}$
	(A) Low	0	0.0001 (0.0030)	0.5	0.4999 (0.0067)	-0.1	-0.1001 (0.0031)	1.5	1.5005 (0.0149)
	(A) High	0	-0.0012 (0.0108)	0.5	0.5007 (0.0217)	-0.1	-0.0985 (0.0111)	1.5	1.5019 (0.0478)
	(B) Mixed	0.41	0.4101 (0.0095)	0.5	0.4999 (0.0064)	-0.58	-0.5753 (0.0113)	-0.5	-0.5036 (0.0057)
p=2		β_{21}	$\hat{\beta}_{21}$	β_{22}	$\hat{\beta}_{22}$	β_{23}	$\hat{\beta}_{23}$	β_{24}	$\hat{\beta}_{24}$
	(A) Low	0	-1.E-05 (2.7E-04)	0	2.9E-06 (2.6E-04)	0	6.9E-06 (2.7E-04)	-0.1	-0.1000 (0.0008)
	(A) High	0	8.5E-06 (9.1E-04)	0	-4.E-05 (8.5E-04)	0	-1.E-05 (8.9E-04)	-0.1	-0.1001 (0.0024)
	(B) Mixed	-0.03	-0.0260 (0.0001)	-0.02	-0.0200 (0.0002)	0.03	0.0328 (0.0008)	0.02	0.0242 (0.0003)
p=3		β_{31}	$\hat{\beta}_{31}$	β_{32}	$\hat{\beta}_{32}$	β_{33}	$\hat{\beta}_{23}$	β_{34}	$\hat{\beta}_{34}$
	(A) Low	–		–		–		–	
	(A) High	–		–		–		–	
	(B) Mixed	4.00E-04	4.0E-04 (2.1E-06)	2.50E-04	2.5E-04 (3.1E-06)	-5.00E-04	-5.E-04 (1.3E-05)	-4.00E-04	-4.E-04 (5.5E-06)
Within-rep Error		σ_1	$\hat{\sigma_1}$	σ_2	$\hat{\sigma_2}$	σ_3	$\hat{\sigma_3}$	σ_4	$\hat{\sigma_4}$
	(A) Low	0.01	0.0033 (0.0006)	0.01	0.0046 (0.0010)	0.01	0.0034 (0.0006)	0.01	0.0065 (0.0020)
	(A) High	0.1	0.0326 (0.0055)	0.1	0.0462 (0.0108)	0.1	0.0330 (0.0055)	0.1	0.0657 (0.0207)
	(B) Mixed	0.01	0.0357 (0.0125)	0.02	0.0959 (0.0417)	0.01	0.0817 (0.0291)	0.03	0.1130 (0.0494)

[1]Low Error (Data with linear or quadratic curves, model fit with p=2 and q=2): σ_k^2=0.01,τ_k^2=0.01; High Error (Data with linear or quadratic curves): σ_k^2=0.1,τ_k^2=0.1; Mixed Error (Data with cubic curves, model fit with q=3 and q=2): $\sigma_k^2 \in (0.01 - 0.03)$, τ_k^2=0.01

Table B.4: Estimated parameters from clustering simulation A data (125 genes per cluster, 4 replicates, 4 clusters) with entropy-penalized EM algorithm Eq2r1 (mixed-effects model with p = 2 and q=2). Only the 4 clusters closest in L2-Norm $|\hat{\beta} - \beta|$ to the true clusters are reported. Mean $\pm$ SD over 1000 iterations.

Parameter	Data	Cluster 1		Cluster 2		Cluster 3		Cluster 4	
Proportions	Error	α_k	$\hat{\alpha_1}$	α_k	$\hat{\alpha_2}$	α_k	$\hat{\alpha_3}$	α_k	$\hat{\alpha_4}$
	(A) Low a	0.25	2.5E-01 (0.0E+00)	0.25	0.2498 (0.0048)	0.25	2.5E-01 (0.0E+00)	0.25	0.2424 (0.0268)
	(A) High	0.25	0.2501 (0.0008)	0.25	0.2491 (0.0107)	0.25	0.2499 (0.0008)	0.25	0.2412 (0.0301)
	(B) Mixed	0.25	0.2234 (0.0517)	0.25	0.2416 (0.0315)	0.25	0.2369 (0.0369)	0.25	0.2455 (0.0230)
Fixed effects									
p=0		β_{01}	$\hat{\beta}_{01}$	β_{02}	$\hat{\beta}_{02}$	β_{03}	$\hat{\beta}_{03}$	β_{04}	$\hat{\beta}_{04}$
	(A) Low	0.5	0.4999 (0.0069)	0.1	0.1003 (0.0068)	0.5	0.5002 (0.0073)	-1	-0.9997 (0.0075)
	(A) High	0.5	0.5002 (0.0225)	0.1	0.0997 (0.0223)	0.5	0.5006 (0.0223)	-1	-1.0009 (0.0242)
	(B) Mixed	-1.05	-1.0492 (0.0345)	-2.3	-2.3011 (0.0601)	2.03	2.0323 (0.0568)	2.85	2.8470 (0.0693)
p=1		β_{11}	$\hat{\beta}_{11}$	β_{12}	$\hat{\beta}_{12}$	β_{13}	$\hat{\beta}_{13}$	β_{14}	$\hat{\beta}_{14}$
	(A) Low	0	3.7E-05 (2.2E-03)	0.5	0.4999 (0.0024)	-0.1	-0.1001 (0.0022)	1.5	1.4996 (0.0062)
	(A) High	0	-0.0002 (0.0072)	0.5	0.5002 (0.0077)	-0.1	-0.1002 (0.0072)	1.5	1.4997 (0.0181)
	(B) Mixed	0.41	0.4103 (0.0061)	0.5	0.4998 (0.0038)	-0.58	-0.5799 (0.0046)	-0.5	-0.5001 (0.0036)
p=2		β_{21}	$\hat{\beta}_{21}$	β_{22}	$\hat{\beta}_{22}$	β_{23}	$\hat{\beta}_{23}$	β_{24}	$\hat{\beta}_{24}$
	(A) Low	0	-5.E-06 (1.9E-04)	0	9.0E-06 (1.9E-04)	0	7.8E-06 (2.0E-04)	-0.1	-0.1000 (0.0004)
	(A) High	0	1.8E-05 (6.4E-04)	0	-4.E-05 (6.2E-04)	0	1.3E-05 (6.4E-04)	-0.1	-0.1000 (0.0012)
	(B) Mixed	-0.03	-0.0260 (0.0001)	-0.02	-0.0200 (0.0001)	0.03	3.3E-02 (9.6E-05)	0.02	0.0240 (0.0002)
p=3		β_{31}	$\hat{\beta}_{31}$	β_{32}	$\hat{\beta}_{32}$	β_{33}	$\hat{\beta}_{23}$	β_{34}	$\hat{\beta}_{34}$
	(A) Low	–		–		–		–	
	(A) High	–		–		–		–	
	(B) Mixed	4.00E-04	4.0E-04 (2.1E-06)	2.50E-04	2.5E-04 (2.6E-06)	-5.00E-04	-5.E-04 (1.9E-06)	-4.00E-04	-4.E-04 (3.5E-06)
Within-rep Error		σ_1	$\hat{\sigma_1}$	σ_2	$\hat{\sigma_2}$	σ_3	$\hat{\sigma_3}$	σ_4	$\hat{\sigma_4}$
	(A) Low	0.01	0.0106 (0.0004)	0.01	0.0115 (0.0003)	0.01	0.0106 (0.0004)	0.01	0.0125 (0.0004)
	(A) High	0.1	0.1054 (0.0038)	0.1	0.1150 (0.0035)	0.1	0.1058 (0.0040)	0.1	0.1245 (0.0042)
	(B) Mixed	0.01	0.0147 (0.0005)	0.02	0.0253 (0.0006)	0.01	0.0153 (0.0005)	0.03	0.0353 (0.0008)

[1]Low Error (Data with linear or quadratic curves, model fit with p=2 and q=2): σ_k^2=0.01,τ_k^2=0.01; High Error (Data with linear or quadratic curves): σ_k^2=0.1,τ_k^2=0.1; Mixed Error (Data with cubic curves, model fit with q=3 and q=2): $\sigma_k^2 \in (0.01 - 0.03)$, τ_k^2=0.01

Table B.4: Estimated parameters from clustering simulation A data (125 genes per cluster, 4 replicates, 4 clusters) with entropy-penalized EM algorithm Eq2r1 (mixed-effects model with p = 2 and q=2). Only the 4 clusters closest in L2-Norm $|\hat{\beta} - \beta|$ to the true clusters are reported. Mean $\pm$ SD over 1000 iterations. (Continued)

Parameter	Data	Cluster 1		Cluster 2		Cluster 3		Cluster 4	
Between-rep Error		τ_k	$\hat{\tau}_1$	τ_k	$\hat{\tau}_2$	τ_k	$\hat{\tau}_3$	τ_k	$\hat{\tau}_4$
	(A) Low	0.01	0.0097 (0.0007)	0.01	0.0096 (0.0007)	0.01	0.0097 (0.0007)	0.01	0.0096 (0.0008)
	(A) High	0.1	0.0968 (0.0071)	0.1	0.0957 (0.0071)	0.1	0.0965 (0.0071)	0.1	0.0953 (0.0084)
	(B) Mixed	0.01	0.0096 (0.0009)	0.01	0.0096 (0.0009)	0.01	0.0096 (0.0008)	0.01	0.0096 (0.0009)
Gene-specific RE		$SD(b_{01})$	$\hat{SD}(b_{01})$	$SD(b_{02})$	$\hat{SD}(b_{02})$	$SD(b_{03})$	$\hat{SD}(b_{03})$	$SD(b_{04})$	$\hat{SD}(b_{04})$
	(A) Low	0.1	0.0247 (0.0066)	0.01	0.0225 (0.0056)	0.03	0.0251 (0.0068)	0.4	0.0242 (0.0060)
	(A) High	0.03	0.0766 (0.0241)	0.04	0.0661 (0.0190)	0.1	0.0764 (0.0238)	0.14	0.0739 (0.0221)
	(B) Mixed	0.21	0.2065 (0.0239)	0.46	0.4578 (0.0418)	0.41	0.4041 (0.0376)	0.57	0.5680 (0.0444)
		$SD(b_{11})$	$\hat{SD}(b_{11})$	$SD(b_{12})$	$\hat{SD}(b_{12})$	$SD(b_{13})$	$\hat{SD}(b_{13})$	$SD(b_{14})$	$\hat{SD}(b_{14})$
	(A) Low	0	0.0071 (0.0019)	0.03	0.0115 (0.0022)	0.01	0.0072 (0.0020)	0.05	0.0276 (0.0040)
	(A) High	0	0.0216 (0.0068)	0.1	0.0360 (0.0070)	0.04	0.0220 (0.0067)	0.17	0.0867 (0.0124)
	(B) Mixed	0.02	0.0217 (0.0043)	0.03	0.0256 (0.0024)	0.03	0.0308 (0.0027)	0.03	0.0262 (0.0022)
		$SD(b_{21})$	$\hat{SD}(b_{21})$	$SD(b_{22})$	$\hat{SD}(b_{22})$	$SD(b_{23})$	$\hat{SD}(b_{23})$	$SD(b_{24})$	$\hat{SD}(b_{24})$
	(A) Low	0	0.0006 (0.0002)	0	0.0006 (0.0001)	0	0.0006 (0.0002)	0.01	0.0017 (0.0003)
	(A) High	0	0.0019 (0.0006)	0	0.0017 (0.0005)	0	0.0019 (0.0006)	0.04	0.0055 (0.0010)
	(B) Mixed	5.20E-04	4.3E-04 (5.1E-05)	4.00E-04	3.0E-04 (3.6E-05)	6.60E-04	5.6E-04 (4.6E-05)	4.80E-04	4.0E-04 (4.6E-05)
		$\rho(b_{0k}, b_{1k})$	$\hat{\rho}(b_{01}, b_{11})$	$\rho(b_{0k}, b_{1k})$	$\hat{\rho}(b_{02}, b_{12})$	$\rho(b_{0k}, b_{1k})$	$\hat{\rho}(b_{03}, b_{13})$	$\rho(b_{0k}, b_{1k})$	$\hat{\rho}(b_{04}, b_{14})$
	(A) Low	-0.5	-0.6049 (0.2326)	-0.5	-0.2878 (0.3037)	-0.5	-0.6080 (0.2290)	-0.5	-0.2936 (0.3097)
	(A) High	-0.5	-0.5972 (0.2720)	-0.5	-0.2807 (0.3198)	-0.5	-0.5998 (0.2811)	-0.5	-0.3047 (0.3275)
	(B) Mixed	-0.7	-0.7133 (0.0930)	-0.7	-0.7004 (0.0838)	-0.7	-0.7005 (0.0785)	-0.7	-0.7026 (0.0579)
		$\rho(b_{0k}, b_{2k})$	$\hat{\rho}(b_{01}, b_{21})$	$\rho(b_{0k}, b_{2k})$	$\hat{\rho}(b_{02}, b_{22})$	$\rho(b_{0k}, b_{2k})$	$\hat{\rho}(b_{03}, b_{23})$	$\rho(b_{0k}, b_{2k})$	$\hat{\rho}(b_{04}, b_{24})$
	(A) Low	0.4	0.5409 (0.2526)	0.4	0.5407 (0.2172)	0.4	0.5346 (0.2561)	0.4	0.2196 (0.3305)
	(A) High	0.4	0.5328 (0.2960)	0.4	0.5417 (0.2360)	0.4	0.5254 (0.3017)	0.4	0.2140 (0.3596)
	(B) Mixed	0.6	0.5476 (0.1380)	0.6	0.5671 (0.1123)	0.6	0.5465 (0.0893)	0.6	0.5455 (0.1009)
		$\rho(b_{1k}, b_{2k})$	$\hat{\rho}(b_{11}, b_{21})$	$\rho(b_{1k}, b_{2k})$	$\hat{\rho}(b_{12}, b_{22})$	$\rho(b_{1k}, b_{2k})$	$\hat{\rho}(b_{13}, b_{23})$	$\rho(b_{1k}, b_{2k})$	$\hat{\rho}(b_{14}, b_{24})$
	(A) Low	-0.9	-0.9720 (0.0204)	-0.9	-0.5054 (0.2415)	-0.9	-0.9526 (0.0389)	-0.9	-0.8897 (0.0620)
	(A) High	-0.9	-0.9712 (0.0277)	-0.9	-0.4886 (0.2699)	-0.9	-0.9495 (0.0492)	-0.9	-0.8885 (0.0720)
	(B) Mixed	-0.8	-0.7812 (0.0855)	-0.8	-0.7804 (0.0588)	-0.8	-0.7830 (0.0485)	-0.8	-0.7738 (0.0604)

[1]Low Error (Data with linear or quadratic curves, model fit with p=2 and q=2): σ_k^2=0.01,τ_k^2=0.01; High Error (Data with linear or quadratic curves): σ_k^2=0.1,τ_k^2=0.1; Mixed Error (Data with cubic curves, model fit with q=3 and q=2): $\sigma_k^2 \in (0.01 - 0.03)$, τ_k^2=0.01

Table B.5: Estimated parameters for each cluster from clustering the bone healing microarray study mice with entropy penalized EM-algorithm (mixed-effects model with p=4, q=2 and strain-(replicate) specific random-effect).

Cluster	$\hat{\alpha}_k$	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	$\hat{\sigma}^2$	$\hat{Var}(b_{0k})$	$\hat{Var}(b_{1k})$	$\hat{Var}(b_{2k})$	$\hat{\rho}(b_{0k}, b_{1k})$	$\hat{\rho}(b_{0k}, b_{2k})$	$\hat{\rho}(b_{1k}, b_{2k})$	$\hat{\tau}_k^2$
1	0.03	-1.87	0.74	-0.061	0.00176	-0.0000173	0.19	0.17	0.0026	0.0000012	-1.00	1.00	-0.12	0.09
2	0.05	-1.84	0.93	-0.090	0.00308	-0.0000354	0.14	0.09	0.0015	0.0000008	-1.00	1.00	-0.13	0.04
3	0.02	-1.07	0.83	-0.093	0.00348	-0.0000429	0.18	0.03	0.0007	0.0000006	-1.00	1.00	-0.16	0.02
4	0.04	-1.18	0.70	-0.070	0.00245	-0.0000285	0.32	0.04	0.0007	0.0000004	-0.99	0.99	-0.13	0.08
5	0.02	-1.60	0.41	-0.017	-0.00007	0.0000067	0.24	0.21	0.0028	0.0000011	-1.00	0.99	-0.11	0.09
6	0.04	-0.36	0.54	-0.068	0.00267	-0.0000341	0.31	0.01	0.0004	0.0000003	-0.99	0.98	-0.17	0.07
7	0.04	-0.53	0.27	-0.023	0.00065	-0.0000061	0.46	0.09	0.0014	0.0000010	-0.93	0.75	-0.12	0.07
8	0.02	-0.88	0.41	-0.035	0.00099	-0.0000095	0.22	0.09	0.0014	0.0000008	-0.97	0.86	-0.12	0.29
9	0.02	-0.62	-0.05	0.027	-0.00156	0.0000240	0.30	0.20	0.0033	0.0000018	-0.97	0.90	-0.13	0.14
10	0.03	0.28	0.23	-0.040	0.00181	-0.0000250	0.44	0.04	0.0004	0.0000003	-0.81	0.32	-0.08	0.06
11	0.06	0.12	0.11	-0.014	0.00053	-0.0000065	0.31	0.04	0.0008	0.0000006	-0.97	0.91	-0.15	0.17
12	0.01	0.25	-0.08	0.006	-0.00022	0.0000027	0.13	0.06	0.0014	0.0000011	-0.96	0.90	-0.15	0.66
13	0.20	0.43	-0.16	0.017	-0.00065	0.0000079	0.34	0.02	0.0004	0.0000003	-0.94	0.83	-0.13	0.08
14	0.11	0.83	-0.39	0.039	-0.00140	0.0000165	0.32	0.01	0.0002	0.0000002	-0.89	0.66	-0.11	0.06
15	0.04	0.86	-0.20	0.008	0.00004	-0.0000032	0.39	0.13	0.0011	0.0000004	-0.95	0.69	-0.08	0.10
16	0.04	0.50	-0.47	0.059	-0.00241	0.0000316	0.37	0.12	0.0026	0.0000016	-0.98	0.94	-0.14	0.13
17	0.03	0.85	-0.41	0.039	-0.00134	0.0000155	0.21	0.07	0.0009	0.0000006	-0.88	0.56	-0.10	0.30
18	0.03	1.31	-0.50	0.037	-0.00096	0.0000082	0.25	0.12	0.0011	0.0000003	-1.00	0.98	-0.10	0.08
19	0.02	1.24	-0.60	0.049	-0.00132	0.0000118	0.16	0.01	0.0002	0.0000001	-0.99	0.99	-0.12	0.02
20	0.04	1.39	-0.88	0.094	-0.00342	0.0000415	0.15	0.01	0.0002	0.0000001	-0.99	0.99	-0.13	0.07
21	0.05	1.66	-0.84	0.082	-0.00276	0.0000316	0.13	0.05	0.0006	0.0000003	-1.00	1.00	-0.11	0.08
22	0.08	1.31	-0.73	0.077	-0.00278	0.0000339	0.26	0.09	0.0013	0.0000006	-0.99	0.96	-0.12	0.10

Table B.6: Estimated parameters for each cluster from clustering the bone healing microarray study mice with modified initialization entropy penalized EM-algorithm (mixed-effects model with p=4, q=2 and strain-(replicate) specific random-effect).

Cluster	$\hat{\alpha}_k$	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	$\hat{\sigma}^2$	$\hat{Var}(b_{0k})$	$\hat{Var}(b_{1k})$	$\hat{Var}(b_{2k})$	$\hat{\rho}(b_{0k}, b_{1k})$	$\hat{\rho}(b_{0k}, b_{2k})$	$\hat{\rho}(b_{1k}, b_{2k})$	$\hat{\tau}_k^2$
1	0.03	-1.91	0.79	-0.067	0.00203	-0.0000209	0.18	0.15	0.0023	0.0000011	-1.00	1.00	-0.12	0.08
2	0.04	-1.80	0.94	-0.092	0.00317	-0.0000367	0.13	0.07	0.0013	0.0000008	-1.00	1.00	-0.14	0.03
3	0.02	-1.06	0.82	-0.092	0.00347	-0.0000429	0.18	0.03	0.0007	0.0000005	-1.00	1.00	-0.17	0.02
4	0.04	-1.17	0.68	-0.068	0.00234	-0.0000271	0.30	0.04	0.0007	0.0000004	-0.99	0.98	-0.13	0.10
5	0.03	-1.57	0.47	-0.027	0.00038	0.0000006	0.23	0.25	0.0037	0.0000016	-1.00	1.00	-0.12	0.12
6	0.04	-0.37	0.55	-0.069	0.00272	-0.0000347	0.31	0.02	0.0004	0.0000003	-0.99	0.98	-0.16	0.06
7	0.04	-0.57	0.30	-0.026	0.00078	-0.0000078	0.46	0.10	0.0016	0.0000010	-0.95	0.81	-0.12	0.07
8	0.03	-0.33	0.26	-0.025	0.00083	-0.0000091	0.25	0.11	0.0020	0.0000013	-0.97	0.89	-0.13	0.29
9	0.02	-0.92	0.05	0.019	-0.00135	0.0000220	0.28	0.20	0.0036	0.0000020	-0.98	0.93	-0.13	0.14
10	0.04	0.26	0.22	-0.037	0.00168	-0.0000231	0.43	0.04	0.0006	0.0000004	-0.83	0.46	-0.10	0.08
11	0.16	0.30	-0.07	0.008	-0.00033	0.0000043	0.35	0.03	0.0005	0.0000003	-0.96	0.85	-0.12	0.09
12	0.02	0.42	-0.15	0.013	-0.00042	0.0000047	0.16	0.06	0.0013	0.0000010	-0.92	0.80	-0.14	0.55
13	0.17	0.68	-0.30	0.031	-0.00111	0.0000130	0.32	0.01	0.0002	0.0000001	-0.90	0.72	-0.12	0.07
14	0.02	-0.01	-0.32	0.050	-0.00227	0.0000314	0.36	0.09	0.0016	0.0000010	-0.94	0.81	-0.13	0.12
15	0.04	0.86	-0.19	0.006	0.00012	-0.0000044	0.38	0.13	0.0011	0.0000004	-0.94	0.66	-0.08	0.10
16	0.08	0.96	-0.54	0.057	-0.00211	0.0000259	0.31	0.08	0.0011	0.0000005	-0.98	0.90	-0.11	0.14
17	0.03	1.29	-0.51	0.041	-0.00114	0.0000108	0.23	0.12	0.0012	0.0000004	-0.99	0.97	-0.10	0.11
18	0.02	1.28	-0.62	0.051	-0.00141	0.0000130	0.16	0.02	0.0002	0.0000001	-1.00	0.99	-0.11	0.02
19	0.07	1.56	-0.87	0.088	-0.00308	0.0000362	0.13	0.05	0.0009	0.0000004	-1.00	1.00	-0.13	0.07
20	0.07	1.39	-0.79	0.084	-0.00306	0.0000373	0.23	0.10	0.0015	0.0000007	-1.00	1.00	-0.12	0.09

Table B.7: Estimated parameters for each cluster from clustering the bone healing microarray study mice with split entropy penalized EM-algorithm (mixed-effects model with p=4, q=2 and strain-(replicate) specific random-effect). Data were clustered into 4 groups of varying polynomial order (p=1 to 4).

Cluster	$\hat{\alpha}_k$	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	$\hat{\sigma}^2$	$\hat{Var}(b_{0k})$	$\hat{Var}(b_{1k})$	$\hat{Var}(b_{2k})$	$\hat{\rho}(b_{0k}, b_{1k})$	$\hat{\rho}(b_{0k}, b_{2k})$	$\hat{\rho}(b_{1k}, b_{2k})$	$\hat{\tau}_k^2$
1	0.03	0.10	-0.004	–	–	–	0.56	0.110	0.0004	–	-1.00	–	–	0.04
2	0.01	0.06	-0.002	–	–	–	0.16	0.022	0.0001	–	-0.97	–	–	0.62
3	0.30	0.17	-0.01	–	–	–	0.37	0.036	0.0001	–	-0.99	–	–	0.13
4	0.03	-0.48	0.10	-0.003	–	–	0.36	0.109	0.0010	0.00000024	-0.99	0.92	-0.09	0.11
5	0.04	0.68	-0.11	0.003	–	–	0.37	0.073	0.0003	0.00000003	-0.96	-0.14	-0.01	0.11
6	0.02	1.17	-0.43	0.027	-0.0004	–	0.20	0.090	0.0013	0.00000065	-0.99	0.99	-0.12	0.18
7	0.01	1.07	-0.48	0.031	-0.0005	–	0.17	0.010	0.0001	0.00000006	-0.99	0.97	-0.11	0.02
8	0.03	-0.75	0.33	-0.022	0.0004	–	0.38	0.126	0.0022	0.00000139	-0.96	0.87	-0.13	0.07
9	0.06	0.75	-0.26	0.018	-0.0003	–	0.30	0.037	0.0007	0.00000053	-0.90	0.74	-0.13	0.05
10	0.02	-1.72	0.56	-0.034	0.0005	–	0.20	0.160	0.0024	0.00000106	-1.00	1.00	-0.12	0.09
11	0.02	-1.18	0.39	-0.024	0.0004	–	0.23	0.334	0.0052	0.00000238	-1.00	0.99	-0.12	0.19
12	0.04	1.02	-0.35	0.022	-0.0004	–	0.32	0.116	0.0013	0.00000057	-0.96	0.84	-0.10	0.11
13	0.06	-0.77	0.67	-0.077	0.0030	-0.00004	0.30	0.076	0.0020	0.00000146	-1.00	1.00	-0.16	0.06
14	0.01	0.20	0.31	-0.056	0.0027	-0.00004	0.37	0.128	0.0031	0.00000296	-0.85	0.65	-0.15	0.04
15	0.08	-1.73	0.89	-0.087	0.0030	-0.00003	0.16	0.171	0.0029	0.00000154	-1.00	1.00	-0.13	0.04
16	0.07	1.36	-0.77	0.082	-0.0030	0.00004	0.24	0.099	0.0016	0.00000078	-1.00	1.00	-0.13	0.10
17	0.08	1.53	-0.85	0.085	-0.0030	0.00003	0.14	0.056	0.0008	0.00000041	-1.00	1.00	-0.12	0.06
18	0.01	-0.97	-0.03	0.034	-0.0020	0.00003	0.26	0.214	0.0038	0.00000253	-0.95	0.84	-0.13	0.05
19	0.06	0.92	-0.56	0.064	-0.0025	0.00003	0.30	0.159	0.0028	0.00000184	-0.95	0.84	-0.13	0.06

Table B.8: Parameter estimates for each cluster from a cluster analysis with the modified initialization entropy penalized EM algorithm with data from all three strains of mice (AJ, B6 and C3H). Mixed-effects model with p=2 and q=2.

Cluster	$\hat{\alpha_k}$	$\hat{\beta}_{0k}$	$\hat{\beta}_{1k}$	$\hat{\beta}_{2k}$	$\hat{\beta}_{3k}$	$\hat{\beta}_{4k}$	$\hat{\sigma}_k^2$	$V(b_{0k})$	$V(b_{1k})$	$V(b_{2k})$	$\rho(b_{0k}, b_{1k})$	$\rho(b_{0k}, b_{2k})$	$\rho(b_{1k}, b_{2k})$
1	0.02	0.15	-2.12	0.91	-0.082	0.00254	-0.0000264	0.17	0.0015	0.00000074	-0.85	0.77	-0.09
2	0.01	0.06	-2.20	0.99	-0.095	0.00335	-0.0000405	0.13	0.0021	0.00000112	-0.97	0.97	-0.13
3	0.03	0.18	-1.73	0.65	-0.048	0.00120	-0.0000100	0.29	0.0031	0.00000166	-0.97	0.99	-0.10
4	0.05	0.16	-1.40	0.85	-0.087	0.00305	-0.0000359	0.10	0.0020	0.00000108	-0.94	0.95	-0.14
5	0.10	0.36	-0.77	0.57	-0.062	0.00225	-0.0000270	0.25	0.0012	0.00000072	-0.75	0.72	-0.07
6	0.04	0.30	-0.90	0.19	0.002	-0.00066	0.0000134	0.46	0.0034	0.00000161	-0.88	0.91	-0.09
7	0.12	0.56	0.27	0.09	-0.015	0.00068	-0.0000092	0.16	0.0007	0.00000049	-0.41	0.27	-0.06
8	0.09	0.28	0.37	-0.04	0.002	-0.00004	0.0000004	0.03	0.0002	0.00000012	-0.45	0.26	-0.07
9	0.09	0.42	0.33	-0.26	0.039	-0.00173	0.0000237	0.23	0.0008	0.00000040	-0.90	0.84	-0.06
10	0.19	0.33	0.61	-0.33	0.032	-0.00111	0.0000128	0.16	0.0003	0.00000028	-0.46	0.30	-0.04
11	0.03	0.18	1.30	-0.66	0.074	-0.00291	0.0000375	0.28	0.0025	0.00000135	-0.88	0.86	-0.09
12	0.03	0.27	1.09	-0.29	0.014	-0.00003	-0.0000037	0.25	0.0017	0.00000042	-0.56	0.38	-0.08
13	0.05	0.53	1.19	-0.61	0.056	-0.00177	0.0000187	0.16	0.0003	0.00000013	-0.76	0.44	-0.03
14	0.02	0.12	1.19	-0.55	0.044	-0.00122	0.0000113	0.11	0.0017	0.00000058	-0.87	0.91	-0.12
15	0.03	0.22	2.02	-0.88	0.087	-0.00305	0.0000359	0.17	0.0019	0.00000057	-0.76	0.70	-0.11
16	0.03	0.21	0.87	-0.76	0.081	-0.00298	0.0000361	0.00	0.0001	0.00000004	-0.82	0.72	-0.12
17	0.03	0.09	1.78	-0.86	0.089	-0.00324	0.0000394	0.06	0.0008	0.00000039	-0.86	0.85	-0.12
18	0.04	0.17	1.51	-0.83	0.077	-0.00242	0.0000254	0.03	0.0001	0.00000007	0.01	0.05	-0.06

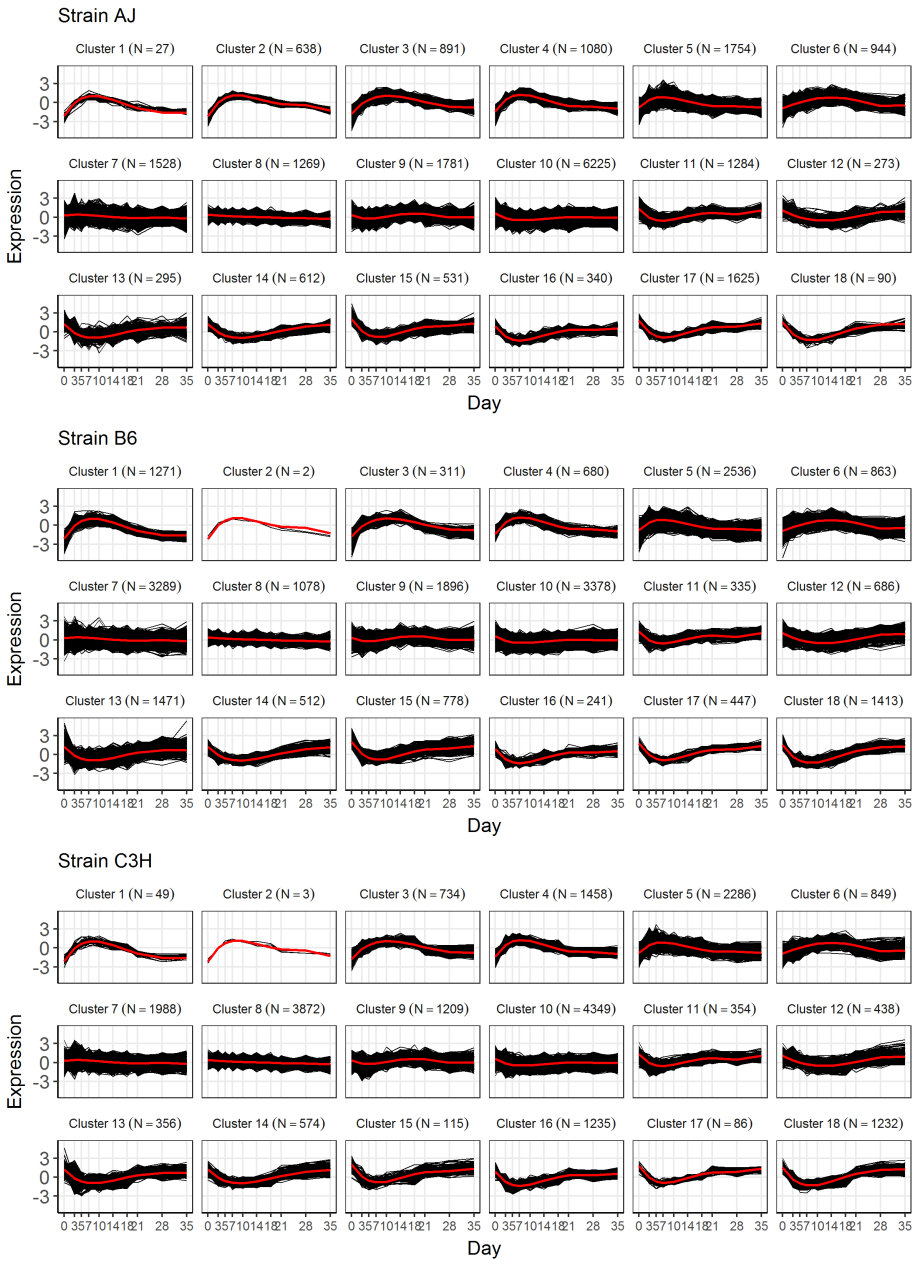

Figure B.1: Strain-specific cluster results from a modified initialization entropy penalized EM algorithm. Each strain-specific set of clusters corresponds to data from 21,187 genes.

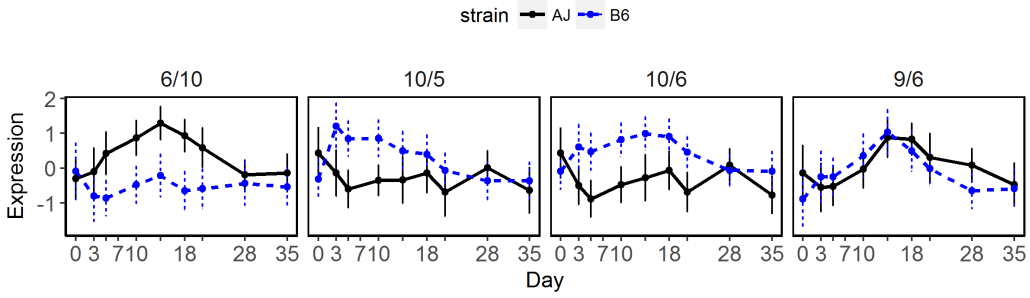

Figure B.2: Genes that were clustered into two different clusters across strain AJ and B6, where one cluster had an initial increasing trend versus a cluster with an initial decreasing trend. Each panel corresponds to genes in Cluster AJ/B6.

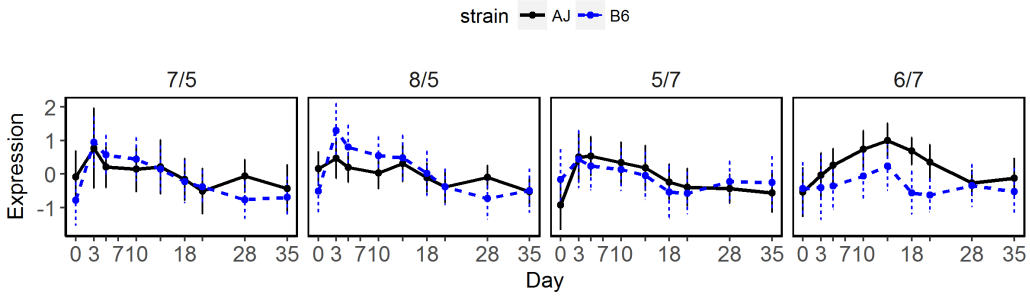

Figure B.3: Genes that were clustered into two different clusters across strain AJ and B6, where one cluster showed an initial increasing trend versus a flat trend. Each panel corresponds to genes in Cluster AJ/B6.

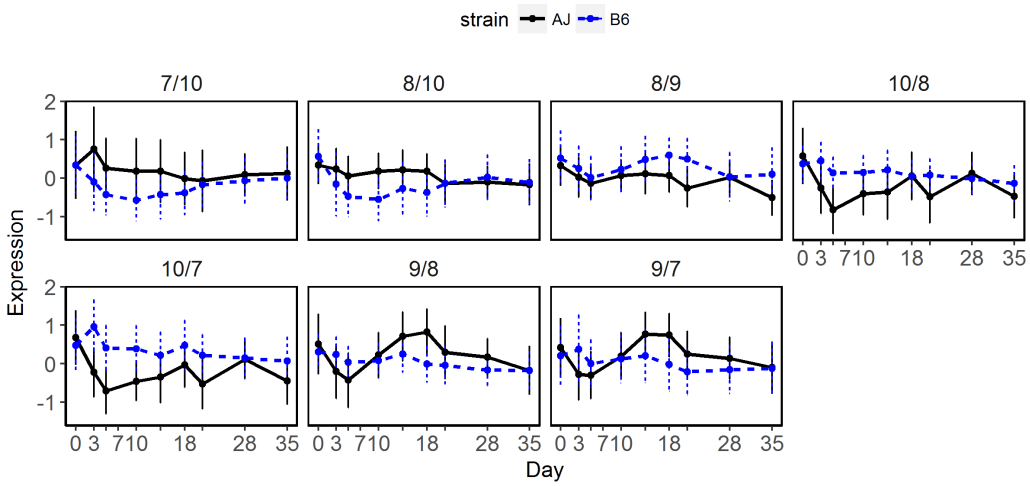

Figure B.4: Genes that were clustered into two different clusters across strain AJ and B6, where one cluster showed an initial decreasing trend versus a flat trend. Each panel corresponds to genes in Cluster AJ/B6.

Table B.9: Enriched KEGG pathways for genes that showed a longer time to maximum expression for AJ mice compared to B6 mice.

Pathway Name	No. Genes
EGFR tyrosine kinase inhibitor resistance	7
Ras signaling pathway	13
Rap1 signaling pathway	14
PI3K-Akt signaling pathway	16
Axon guidance	12
Focal adhesion	13
Adherens junction	9
Signaling pathways regulating pluripotency of stem cells	10
Pathways in cancer	21
MicroRNAs in cancer	13
Prostate cancer	8
Basal cell carcinoma	5
Fluid shear stress and atherosclerosis	10

Table B.10: Enriched KEGG pathways for genes that showed a longer time to minimum expression for AJ mice compared to B6 mice.

Pathway Name	No. Genes
Pyrimidine metabolism	17
Lysine degradation	12
Selenocompound metabolism	5
Inositol phosphate metabolism	12
Glycosylphosphatidylinositol (GPI)-anchor biosynthesis	6
Aminoacyl-tRNA biosynthesis	14
RNA transport	34
mRNA surveillance pathway	23
RNA degradation	23
Basal transcription factors	13
DNA replication	7
Spliceosome	41
Base excision repair	9
Mismatch repair	6
Homologous recombination	21
Fanconi anemia pathway	21
NF-kappa B signaling pathway	19
FoxO signaling pathway	32
Phosphatidylinositol signaling system	16
Cell cycle	41
Oocyte meiosis	20
p53 signaling pathway	16
Ubiquitin mediated proteolysis	26
Mitophagy - animal	15

Pathway Name	No. Genes
AMPK signaling pathway	19
Longevity regulating pathway	18
Cellular senescence	38
Hematopoietic cell lineage	18
IL-17 signaling pathway	15
Th1 and Th2 cell differentiation	14
Th17 cell differentiation	16
B cell receptor signaling pathway	22
Fc epsilon RI signaling pathway	13
Progesterone-mediated oocyte maturation	16
Adipocytokine signaling pathway	12
Type II diabetes mellitus	9
Hepatitis B	23
Influenza A	28
Human T-cell leukemia virus 1 infection	38
Herpes simplex infection	41
Epstein-Barr virus infection	38
Viral carcinogenesis	35
Pancreatic cancer	13
Prostate cancer	17
Chronic myeloid leukemia	14
Small cell lung cancer	19
Primary immunodeficiency	9

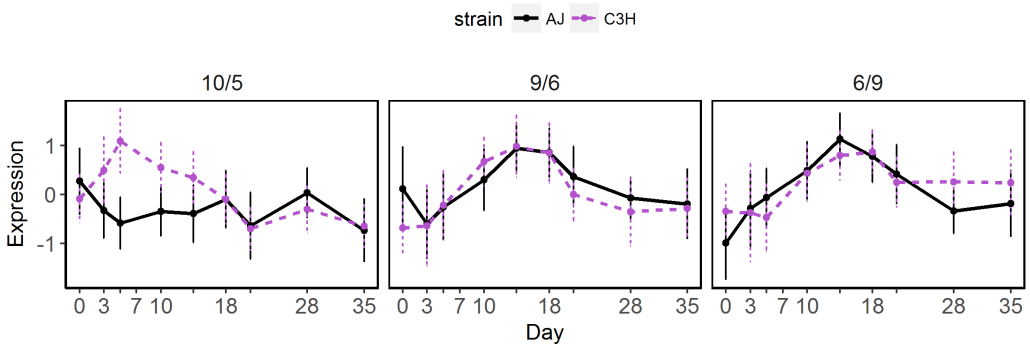

Figure B.5: Genes that were clustered into two different clusters across strain AJ and C3H, where one cluster had an initial increasing trend versus a cluster with an initial decreasing trend.

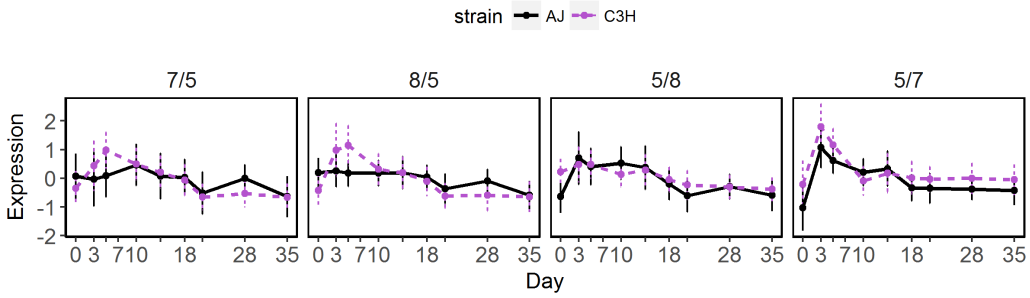

Figure B.6: Genes that were clustered into two different clusters across strain AJ and C3H, where one cluster showed an initial increasing trend versus a flat trend (horizontal).

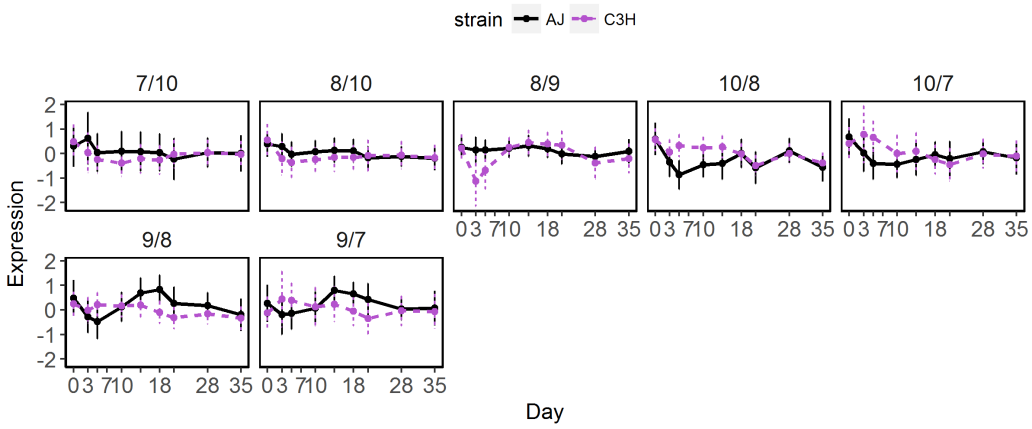

Figure B.7: Genes that were clustered into two different clusters across strain AJ and C3H, where one cluster showed an initial decreasing trend versus a flat trend (horizontal).

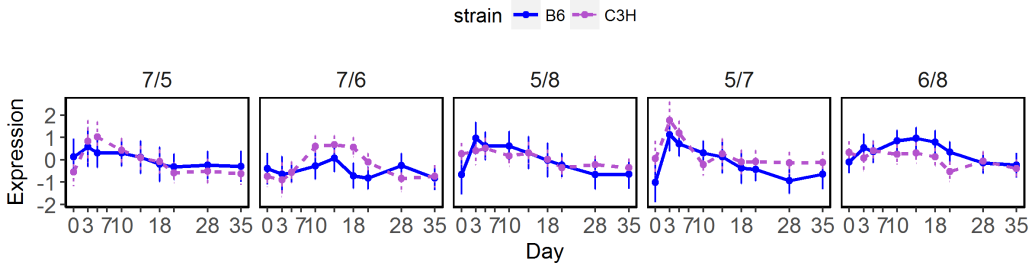

Figure B.8: Genes that were clustered into two different clusters across strain B6 and C3H, where one cluster showed an initial increasing trend versus a flat trend (horizontal).

Table B.11: Enriched KEGG pathways for genes that had a longer time to maximum expression in AJ mice compared to C3H mice.

Pathway Name	No. Genes
Arginine and proline metabolism	10
N-Glycan biosynthesis	12
Other types of O-glycan biosynthesis	5
Mannose type O-glycan biosynthesis	8
Glycosaminoglycan biosynthesis - chondroitin sulfate / dermatan sulfate	5
Glycosaminoglycan biosynthesis - heparan sulfate / heparin	7
Inositol phosphate metabolism	13
Metabolic pathways	110
Rap1 signaling pathway	22
HIF-1 signaling pathway	13
Phosphatidylinositol signaling system	15
Protein processing in endoplasmic reticulum	32
mTOR signaling pathway	17
PI3K-Akt signaling pathway	43
Wnt signaling pathway	20
Hedgehog signaling pathway	7
Axon guidance	29
Hippo signaling pathway	28
Hippo signaling pathway - multiple species	13
Focal adhesion	34

Pathway Name	No. Genes
ECM-receptor interaction	23
Signaling pathways regulating pluripotency of stem cells	17
Parathyroid hormone synthesis, secretion and action	17
AGE-RAGE signaling pathway in diabetic complications	18
Protein digestion and absorption	13
Cocaine addiction	8
Amoebiasis	16
Human papillomavirus infection	42
Pathways in cancer	56
Proteoglycans in cancer	29
Pancreatic cancer	12
Basal cell carcinoma	15
Small cell lung cancer	15
Non-small cell lung cancer	10
Breast cancer	18
Hepatocellular carcinoma	20
Gastric cancer	19
Hypertrophic cardiomyopathy (HCM)	12
Arrhythmogenic right ventricular cardiomyopathy (ARVC)	11
Dilated cardiomyopathy (DCM)	13

Table B.12: Enriched KEGG pathways for genes that had a longer time to minimum expression in AJ mice compared to C3H mice.

Pathway Name	No. Genes
Pyrimidine metabolism	21
Lysine degradation	13
Inositol phosphate metabolism	16
Platinum drug resistance	20
ABC transporters	11
DNA replication	24
Spliceosome	28
Nucleotide excision repair	15
Mismatch repair	14
Homologous recombination	18
Fanconi anemia pathway	21
FoxO signaling pathway	26
Phosphatidylinositol signaling system	19
Cell cycle	55
Oocyte meiosis	21
p53 signaling pathway	23
Mitophagy - animal	13
Apoptosis - multiple species	9
Cellular senescence	34
Platelet activation	25
Hematopoietic cell lineage	26
B cell receptor signaling pathway	17
Fc epsilon RI signaling pathway	15
Progesterone-mediated oocyte maturation	18
Alcoholism	32
Viral carcinogenesis	45
Asthma	7
Systemic lupus erythematosus	30

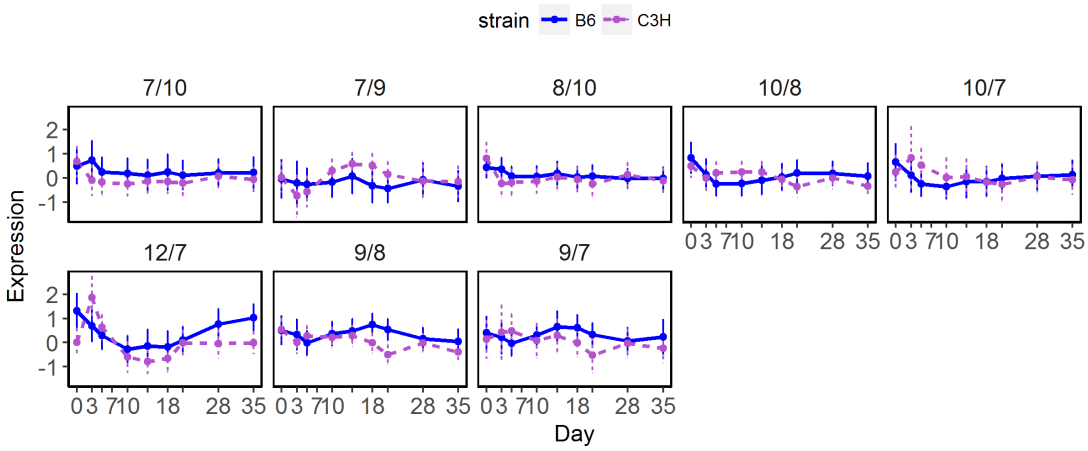

Figure B.9: Genes that were clustered into two different clusters across strain B6 and C3H, where one cluster showed an initial decreasing trend versus a flat trend (horizontal).

Table B.13: Enriched KEGG pathways for genes that had a longer time to maximum expression in B6 mice compared to C3H mice.

Pathway Name	No. Genes
Fructose and mannose metabolism	6
Lysine degradation	8
Arginine and proline metabolism	10
N-Glycan biosynthesis	10
Other types of O-glycan biosynthesis	6
Mannose type O-glycan biosynthesis	7
Amino sugar and nucleotide sugar metabolism	8
Glycosaminoglycan biosynthesis - chondroitin sulfate / dermatan sulfate	7
Glycosaminoglycan biosynthesis - heparan sulfate / heparin	6
Inositol phosphate metabolism	11
Metabolic pathways	106
EGFR tyrosine kinase inhibitor resistance	11
Endocrine resistance	13
MAPK signaling pathway	28
ErbB signaling pathway	12
Ras signaling pathway	26
Rap1 signaling pathway	32
Calcium signaling pathway	21
cGMP-PKG signaling pathway	19
HIF-1 signaling pathway	20
Phospholipase D signaling pathway	20
Protein processing in endoplasmic reticulum	28
Endocytosis	28
PI3K-Akt signaling pathway	48
Wnt signaling pathway	23
Hedgehog signaling pathway	7
Axon guidance	33
VEGF signaling pathway	8
Apelin signaling pathway	16
Hippo signaling pathway	27
Hippo signaling pathway - multiple species	10
Focal adhesion	39
ECM-receptor interaction	25
Tight junction	18
Gap junction	12
Signaling pathways regulating pluripotency of stem cells	21
Circadian entrainment	13
Glutamatergic synapse	15
Cholinergic synapse	15
Regulation of actin cytoskeleton	27
Insulin secretion	15
GnRH signaling pathway	12
Melanogenesis	15
Oxytocin signaling pathway	17
Glucagon signaling pathway	12
Aldosterone synthesis and secretion	13
Relaxin signaling pathway	21
Cortisol synthesis and secretion	10
Parathyroid hormone synthesis, secretion and action	17
AGE-RAGE signaling pathway in diabetic complications	20
Cushing syndrome	21
Protein digestion and absorption	12
Bacterial invasion of epithelial cells	10
Malaria	7
Toxoplasmosis	13
Amoebiasis	21
Human papillomavirus infection	42
Human T-cell leukemia virus 1 infection	27
Pathways in cancer	72
Proteoglycans in cancer	34
Renal cell carcinoma	11
Pancreatic cancer	13
Endometrial cancer	9
Basal cell carcinoma	13
Chronic myeloid leukemia	10
Small cell lung cancer	19
Non-small cell lung cancer	10
Breast cancer	20
Hepatocellular carcinoma	21
Gastric cancer	22
Central carbon metabolism in cancer	9
Hypertrophic cardiomyopathy (HCM)	14
Arrhythmogenic right ventricular cardiomyopathy (ARVC)	14
Dilated cardiomyopathy (DCM)	17
Fluid shear stress and atherosclerosis	16

Table B.14: Enriched KEGG pathways for genes that had a longer time to minimum expression in B6 mice compared to C3H mice.

Pathway Name	No. Genes
Lysine degradation	9
mRNA surveillance pathway	12
Homologous recombination	10
Fanconi anemia pathway	9
Cell cycle	17
p53 signaling pathway	11
Ubiquitin mediated proteolysis	15
NOD-like receptor signaling pathway	19
RIG-I-like receptor signaling pathway	10

BIBLIOGRAPHY

Aike, H. A. I. (1974). A New Look at the Statistical Model Identification. *IEEE Transactions on Automatic Control*, *19*(6), 716–723.

Banfield, J. D., & Raftery, A. E. (1993). Model-based gaussian and non-gaussian clustering. *Biometrics*, (pp. 803–821).

Biernacki, C., & Celeux, G. (2003). Choosing starting values for the EM algorithm for getting the highest likelihood in multivariate Gaussian mixture models. *Computational Statistics & Data Analysis*, *41*, 561–575.

Biernacki, C., Celeux, G., Biernacki, C., & Celeux, G. (2000). Assessing a Mixture Model for Clustering with the Integrated Classification Likelihood. *IEEE Transactions on pattern analysis and machine intelligence*, *22*(7), 719–725.

Brivanlou, A. H., & Darnell, J. E. (2002). Signal transduction and the control of gene expression. *Science*, *295*(5556), 813–818.

Burnham, K. P., & Anderson, D. R. (2003). *Model selection and multimodel inference: a practical information-theoretic approach*. Springer Science & Business Media.

Celeux, G., Martin, O., & Lavergne, C. (2005). Mixture of linear mixed models for clustering gene expression profiles from repeated microarray experiments. *Statistical Modeling*, *5*, 243–267.

Chamroukhi, F. (2015). Unsupervised learning of regression mixture models with unknown number of components. *Journal of Statistical Computation and Simulation*, *86*(12), 2308–2334.

Chen, M.-L., Shah, V., Patnaik, R., Adams, W., Hussain, A., Conner, D., Mehta, M., Malinowski, H., Lazor, J., Huang, S.-M., et al. (2001). Bioavailability and bioequivalence: an fda regulatory overview. *Pharmaceutical research*, *18*(12), 1645–1650.

Dempster, A. A. P., Laird, N. M., Rubin, D. B., Journal, S., Statistical, R., & Series, S. (1977). Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society*, *39*(1), 1–38.

Eisen, M. B., Spellman, P. T., Brown, P. O., & Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences*, *95*(25), 14863–14868.

Figueiredo, M. A. T., Member, S., & Jain, A. K. (2002). Unsupervised Learning of Finite Mixture Models. *IEEE Transactions on pattern analysis and machine intelligence*, *24*(3), 381–396.

Fonesca, J. R., & G.M.S, C. M. (2007). Mixture-model cluster analysis using information theoretical criteria. *Intelligent Data Analysis*, *11*, 155–173.

Food, Administration, D., et al. (2014). Guidance for industry: bioavailability and bioequivalence studies submitted in ndas or inds–general considerations. *Rockville, MD: Food and Drug Administration*.

Fraley, C., Raftery, A. E., Murphy, T. B., & Scrucca, L. (2012). mclust Version 4 for R : Normal Mixture Modeling for Model-Based Clustering , Classification , and Density Estimation. Tech. rep.

Friedman, J., Hastie, T., & Tibshirani, R. (2001). *The elements of statistical learning*, vol. 1. Springer series in statistics New York.

Gaffney, S. (2004). *Probabilistic curve-aligned clustering and prediction with regression mixture models*. Ph.D. thesis, University of California, Irvine.

Gaffney, S., & Smyth, P. (1999). Trajectory Clustering with Mixtures of Regression Models. In *Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*, (pp. 63–72).

Ghandhi, S. A., Sinha, A., Markatou, M., & Amundson, S. A. (2011). Time-series clustering of gene expression in irradiated and bystander fibroblasts: an application of fbpa clustering. *BMC genomics*, *12*(1), 2.

Glass, D. A., Bialek, P., Ahn, J. D., Starbuck, M., Patel, M. S., Clevers, H., Taketo, M. M., Long, F., McMahon, A. P., Lang, R. A., et al. (2005). Canonical wnt signaling in differentiated osteoblasts controls osteoclast differentiation. *Developmental cell*, *8*(5), 751–764.

Grimes, R., Jepsen, K. J., Fitch, J. L., Einhorn, T. A., & Gerstenfeld, L. C. (2011). The transcriptome of fracture healing defines mechanisms of coordination of skeletal and vascular development during endochondral bone formation. *Journal of Bone and Mineral Research*, *26*(11), 2597–2609.

Hartigan, J. A., & Wong, M. A. (1979). Algorithm AS 136 A K-Means Clustering Algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, *28*(1), 100–108.

Heinze, G., Wallisch, C., & Dunkler, D. (2018). Variable selection–a review and recommendations for the practicing statistician. *Biometrical Journal*, *60*(3), 431–449.

Hubert, L., & Arabie, P. (1985). Comparing partitions. *Journal of Classification*, *2*(1), 193–218.

Irizarry, R. A., Hobbs, B., Collin, F., Beazer-Barclay, Y. D., Antonellis, K. J., Scherf, U., & Speed, T. P. (2003). Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, *4*(2), 249–264.

Jain, A. K., Murty, M. N., & Flynn, P. J. (1999). Data clustering: a review. *ACM computing surveys (CSUR)*, *31*(3), 264–323.

Jepsen, K. J., Price, C., Silkman, L. J., Nicholls, F. H., Nasser, P., Hu, B., Hadi, N., Alapatt, M., Stapleton, S. N., Kakar, S., Einhorn, T. A., & Gerstenfeld, L. C. (2008). Genetic variation in the patterns of skeletal progenitor cell differentiation and progression during endochondral bone formation affects the rate of fracture healing. *Journal of Bone and Mineral Research*, *23*(8), 1204–1216.

Kaufman, L., & Rousseeuw, P. J. (2009). *Finding groups in data: an introduction to cluster analysis*, vol. 344. John Wiley & Sons.

Lakshmi, R. V., & Vaidyanathan, V. (2016). Parameter estimation in gamma mixture model using normal-based approximation. *Journal of Statistical Theory and Applications*, *15*(1), 25–35.

Lindsay, B. G. (1986). Exponential family mixture models (with least-squares estimators). *The Annals of Statistics*, *14*(1), 124–137.

Lu, D., Tripodis, Y., Gerstenfeld, L., Demissie, S., & Wren, J. (2018). Clustering of temporal gene expression data with mixtures of mixed effects models with a penalized likelihood. *Bioinformatics*.

McLachlan, G. J., Bean, R., & Peel, D. (2002). A mixture model-based approach to the clustering of microarray expression data. *Bioinformatics*, *18*(3), 413–422.

McLachlan, G. J., & Peel, D. (2000). Assessing the number of components in mixture models. In *Finite Mixture Models*, (pp. 175–220).

Ng, S. K., Mclachlan, G. J., Wang, K., Jones, L. B.-t., & Ng, S. (2006). A mixture model with random-effects components for clustering correlated gene-expression profiles. *Biostatistics*, *22*(14), 1745–1752.

NIH, C. (2000). Osteoporosis prevention, diagnosis, and therapy. *NIH consensus statement*, *17*, 1–45.

Pedraza, J. M., & Paulsson, J. (2007). Random timing in signaling cascades. *Molecular systems biology*, *3*(1), 81.

Praemer, A., Furner, S., Rice, D. P., & Kelsey, J. L. (1992). Musculoskeletal conditions in the united states.

Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical association*, *66*(336), 846–850.

Roeder, K., Wasserman, L., Roeder, K., & Wasserman, L. (1997). Practical Bayesian Density Estimation Using Mixtures of Normals Stable URL : http://www.jstor.org/stable/2965553 Practical Bayesian Density Estimation Using Mixtures of Normals. *Journal of the American Statistical Association*, *92*(439), 894–902.

Rowicka, M., Kudlicki, A., Tu, B. P., & Otwinowski, Z. (2007). High-resolution timing of cell cycle-regulated gene expression. *Proceedings of the National Academy of Sciences*, *104*(43), 16892–16897.

Schwarz, G., et al. (1978). Estimating the dimension of a model. *The annals of statistics*, *6*(2), 461–464.

Shao, J. (1993). Linear model selection by cross-validation. *Journal of the American Statistical Association*, *88*(422), 486–494.

Singh, A., & Dennehy, J. J. (2014). Stochastic holin expression can account for lysis time variation in the bacteriophage λ. *Journal of The Royal Society Interface*, *11*(95), 20140140.

Sneath, P. H., & Sokal, R. R. S. (1973). *Numerical taxonomy. The principles and practice of numerical classification.*. San Francisco, 1 ed.

Stone, M. (1977). An asymptotic equivalence of choice of model by cross-validation and akaike's criterion. *Journal of the Royal Statistical Society. Series B (Methodological)*, (pp. 44–47).

Storey, J. D., Xiao, W., Leek, J. T., Tompkins, R. G., & Davis, R. W. (2005). Significance analysis of time course microarray experiments. *Pnas*, *102*(36), 12837–12842.

Thouverey, C., & Caverzasio, J. (2015). Focus on the p38 mapk signaling pathway in bone development and maintenance. *BoneKEy reports*, *4*.

US, D. (2004). Bone health and osteoporosis: a report of the surgeon general. *http://www. surgeongeneral. gov/library/bonehealth/content. html*.

Wall, M., Rechtsteiner, A., & Rocha, L. (2003). Singular value decomposition and principal component analysis. *A Practical Approach to Microarray Data Analysis*, (pp. 91–109).

Wigner, N. A., Luderer, H. F., Cox, M. K., Sooy, K., Gerstenfeld, L. C., & Demay, M. B. (2010). Acute phosphate restriction leads to impaired fracture healing and resistance to bmp-2. *Journal of Bone and Mineral Research*, *25*(4), 724–733.

Yang, M.-s., Lai, C.-y., & Lin, C.-y. (2012). A robust EM clustering algorithm for Gaussian mixture models. *Pattern Recognition*, *45*(11), 3950–3961.

Yeung, K. Y., Fraley, C., Murua, A., Raftery, A. E., & Ruzzo, W. L. (2001). Model-based clustering and data transformations for gene expression data. *Bioinformatics*, *17*(10), 977–987.

Zhao, S., Guo, Y., & Shyr, Y. (2015). *KEGGprofile: An annotation and visualization package for multi-types and multi-groups expression data in KEGG pathway*. R package version 1.16.0.

Zou, W., Izawa, T., Zhu, T., Chappel, J., Otero, K., Monkley, S. J., Critchley, D. R., Petrich, B. G., Morozov, A., Ginsberg, M. H., et al. (2013). Talin1 and rap1 are critical for osteoclast function. *Molecular and cellular biology*, *33*(4), 830–844.

CURRICULUM VITAE